# ULTRA LOW POWER NON-BOOLEAN COMPUTING WITH TUNNELING FIELD-EFFECT-TRANSISTORS

A Dissertation
Presented to
The Academic Faculty

By

Amit Ranjan Trivedi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2015

# ULTRA LOW POWER NON-BOOLEAN COMPUTING WITH TUNNELING FIELD-EFFECT-TRANSISTORS

Approved by:


Dr. Saibal Mukhopadhyay, Advisor
*Associate Professor, School of ECE*
*Georgia Institute of Technology*

Dr. Sudhakar Yalamanchili
*Professor, School of ECE*
*Georgia Institute of Technology*

Dr. Arijit Raychowdhury
*Associate Professor, School of ECE*
*Georgia Institute of Technology*

Dr. Hyesoon Kim
*Associate Professor, College of Computing*
*Georgia Institute of Technology*

Dr. Azad Naeemi
*Assistant Professor, School of ECE*
*Georgia Institute of Technology*



Date Approved: September 2015

# TABLE OF CONTENTS

**LIST OF TABLES** . . . . . vi

**LIST OF FIGURES** . . . . . vii

**CHAPTER 1 INTRODUCTION** . . . . . 1
1.1 A Paradigm-Shift in the Contemporary Computing: Added Emphasis to Energy-Efficiency . . . . . 1
1.2 Energy-efficient Computing Approach of the Dissertation . . . . . 2
1.2.1 Explore the Role of Emerging Transistor Technologies . . . . . 2
1.2.2 Explore the Role of Application-Specific and Intelligent Hardware 3
1.3 Contributions of the Dissertation . . . . . 3
1.4 Organization of the Dissertation . . . . . 5

**CHAPTER 2 LITERATURE SURVEY** . . . . . 6
2.1 Switching-Slope-induced Energy-Efficiency Limitations in MOSFET-based Computing . . . . . 6
2.1.1 Limitations in MOSFET-based Digital-Computing . . . . . 7
2.1.2 Limitations in MOSFET-based Analog-Computing . . . . . 7
2.1.3 Emerging Transistor Technologies with Steeper Switching-Slope . 7
2.2 Tunneling-Field-Effect Transistors (Tunnel-FET or TFET) . . . . . 8
2.2.1 Charge-Conduction in TFET . . . . . 9
2.2.2 Silicon-channel TFET . . . . . 11
2.2.3 Alternate Channel-Materials for TFET . . . . . 12
2.3 Non-Boolean Computing Architecture . . . . . 12
2.3.1 Cellular Neural Network (CNN) . . . . . 13
2.3.1.1 CNN-based Image Processing . . . . . 15
2.3.1.2 CNN-based Associative Memory . . . . . 16
2.3.2 Grid-based Non-Boolean Associative Computing with Emerging Devices . . . . . 17
2.4 Conclusions . . . . . 18

**CHAPTER 3 EMERGING DEVICE CHARACTERISTICS VARIABILITY MECHANISMS IN NON-TRADITIONAL MOSFET DESIGN AND INTEGRATION** . . . . . 20
3.1 Threshold-Voltage Variability in Non-traditional High-$\kappa$ Dielectric MOSFET . . . . . 20
3.1.1 Oxygen-vacancy Generation Model . . . . . 21
3.1.2 Simulation Methodology for Oxygen Vacancy-induced Variability 23
3.1.3 Oxygen Vacancy-induced Variability in FinFET . . . . . 25
3.1.4 Implications of Metal Grain Granularity to Oxygen Vacancy Distribution . . . . . 27

3.2 Threshold Voltage Modulation in Non-traditional Three-Dimensional MOSFET Integration . . . 29
3.2.1 Through-Oxide-Via-induced Back-Gate Effect . . . 29
3.2.2 Through-Oxide-Via-induced On/Off-current and Threshold Voltage Modulation . . . 29
3.3 Conclusions . . . 31

**CHAPTER 4 SINGLE NEURON-BASED POWER-GATING EFFICIENCY LEARNING, AND APPLICATION TO SELF-ADAPTIVE POWER-GATING** 33
4.1 Fine-Grained Power-Gating . . . 33
4.2 Single Neuron Dynamics-based Power-Gating-Efficiency Learner . . . 34
4.3 Power-Gating-Efficiency Learner Implementation and Operation . . . 34
4.3.1 Power-Gating-Efficiency Learner Implementation . . . 34
4.3.2 Break-Even and Power-Gating-Efficiency Tracking . . . 36
4.4 Self-Adaptive Power-Gating with Power-Gating-Efficiency Learner . . . 38
4.5 Test-chip and Measurement Results . . . 38
4.5.1 Characterization of Break-Even Learning Accuracy . . . 40
4.5.2 Self-Adaptive Power-Gating at varying Process/Temperature Conditions . . . 41
4.5.3 Implications of the Learning Cycle . . . 42
4.6 Limitations of the PGE Learner-based Power-Gating . . . 44
4.7 Comparison with the Prior Works . . . 46
4.8 Conclusions . . . 46

**CHAPTER 5 TUNNELING-FIELD-EFFECT-TRANSISTORS TO ENABLE ULTRA-LOW-POWER ANALOG-COMPUTING** . . . 48
5.1 Motivation to Employ TFET for Low-Power Analog-Computing . . . 48
5.2 TFET Characteristics . . . 50
5.2.1 Simulations of TFET . . . 50
5.2.2 Calibration of TFET Simulation Parameters . . . 50
5.2.3 Electrical-Characteristics of TFETs of Various Channel Materials . 54
5.3 TFET-based Circuit Simulation Methodology . . . 58
5.4 TFET-based Operational-Transconductance-Amplifier . . . 58
5.5 Conclusions . . . 62

**CHAPTER 6 TFET CELLULAR NEURAL NETWORK-BASED IMAGE PROCESSING** . . . 64
6.1 Cellular Neural Network Cell Design Constraints . . . 64
6.2 Cellular Neural Network Cell Power Scaling . . . 65
6.2.1 CNN Cell Power Scaling Approach . . . 65
6.2.2 CNN Cell Power Scaling Simulation Results . . . 65
6.3 Simulation Methodology for CNN-based Image Processing . . . 67
6.4 Comparison of TFET and FinFET CNN-based Image Processing . . . 68
6.4.1 Fixed-Array Approach . . . 68

6.4.2 Energy-optimal Approach . . . . . 70
6.5 Conclusions . . . . . 71

**CHAPTER 7 TFET CELLULAR NEURAL NETWORK-BASED ASSOCIATIVE MEMORY** . . . . . 72
7.1 Motivation for CNN Associative Memory with TFET . . . . . 72
7.2 TFET-based CNN-AM Simulation Methodology . . . . . 74
7.3 TFET-based CNN-AM Simulation Results . . . . . 74
7.3.1 CNN-AM at NR of One . . . . . 74
7.3.2 Improving TFET-CNN-AM by High NR Design . . . . . 77
7.4 CNN-AM under Process Variations . . . . . 80
7.5 Implementation Complexity in a Higher NR TFET-CNN . . . . . 81
7.6 TFET-CNN for Image Processing vs. Associative Memory . . . . . 82
7.7 Conclusions . . . . . 83

**CHAPTER 8 NON-CONVENTIONAL III-V HETEROJUNCTION TFET FOR AREA AND ENERGY-EFFICIENT NEUROMORPHIC ASSOCIATIVE PROCESSING** . . . . . 84
8.1 Distance-Computation-based Neuromorphic Architecture for Associative-Processing . . . . . 84
8.2 Gate/Source-overlapped Heterojunction-TFET as a Single-Transistor Distance-Computing-Cell . . . . . 85
8.3 Programming of SO-HTFET Distance-Computing-Cell . . . . . 89
8.4 SO-HTFET-based Associative Processing . . . . . 93
8.5 HTFET Winner-Take-All . . . . . 94
8.6 On-line Training and Plasticity in SO-HTFET-AP . . . . . 95
8.7 Face-recognition using SO-HTFET Associative Processing Platform . . . 97
8.8 Reinforced learning in SO-HTFET Associative Processing . . . . . 100
8.9 Comparison with the Existing Approaches . . . . . 101
8.10 Conclusions . . . . . 102

**CHAPTER 9 SUMMARY AND FUTURE WORKS** . . . . . 105

**REFERENCES** . . . . . 108

# LIST OF TABLES


Table 1 Proto-type chip specifications . . . . . 39

Table 2 Comparison of PGE-learner with the prior-art . . . . . 46

Table 3 Si-Ge TFET simulation parameters . . . . . 52

Table 4 III-V TFET simulation parameters . . . . . 53

Table 5 TFET geometrical and process specifications . . . . . 54

Table 6 TFET-OTA design specifications . . . . . 60

Table 7 Direct-tunneling parameters for various materials . . . . . 91

Table 8 Comparison of SO-HTFET-AP to the prior-works . . . . . 104

# LIST OF FIGURES


Figure 1 Near future growth-rate of various computing-platforms. . . . . 1

Figure 2 (a) TFET schematic. (b) Energy-band diagram and band-to-band-tunneling in TFET. . . . . 10

Figure 3 Silicon-channel and non-silicon-channel approaches to improve on-current of TFET. . . . . 11

Figure 4 (a) Cellular neural network (CNN) schematic. (b) CNN neuron interconnections for the neighborhood-radius of two. . . . . 14

Figure 5 An analog implementation of cellular neural network. . . . . 15

Figure 6 Cellular neural network-based image-processing: edge-detection and noise-filtering. . . . . 16

Figure 7 Cellular neural network-based associative-memory operation. . . . . 17

Figure 8 A Hebbian learning-based hetero-associative recall algorithm. . . . . 17

Figure 9 A demonstration of oxygen-vacancy generation in $HfO_2$. . . . . 21

Figure 10 (a) Surplus electrons fall from the trap energy level into the lower energy metal electrode, and OV dipole forms. (b) OV induced dipole created by the mirror charge. . . . . 22

Figure 11 (a) FinFET schematic. (b) Charged cubes (0.5 nm) representing a random placement and count of positively charged OV in the $HfO_2$ layer to induce local variability. . . . . 24

Figure 12 OV-induced local variability in surface potential at the channel/oxide interface. . . . . 25

Figure 13 (a) OV-induced $V_{th}$ shift and variability in $I_{DS}$-$V_{GS}$ characteristics of 22-nm FinFET. (b) Histogram of $V_{th}$ variability in 22-nm FinFET. (c) Correlation of OV-induced $V_{th}$ quantiles against the standard normal across technologies. . . . . 26

Figure 14 OV-induced variability compared with the other sources of variability: RDF, MGG, FER, and GER. . . . . 27

Figure 15 (a) Under metal gate granularity, a higher oxygen vacancy concentration under higher workfunction metal grain. (b) MGG-induced variability suppression at varying oxygen vacancy concentration. . . . . 28

Figure 16 Interaction of TOV and FD-SOI device in a 3-D integrated system. . . . . 29

Figure 17 Effect of the TOV on the NFET: (a) threshold voltage, (b) subthreshold leakage $I_{off}$ , and (c) on-current ($I_{on}$). [TOV diameter, blue: 1 $\mu$m, black: 2 $\mu$m, red: 5 $\mu$m ] . . . . . 31

Figure 18 Effect of the TOV on the PFET: (a) threshold voltage, (b) subthreshold leakage $I_{off}$ , and (c) on-current ($I_{on}$). [TOV diameter, blue: 1 $\mu$m, black: 2 $\mu$m, red: 5 $\mu$m ] . . . . . 31

Figure 19 A demonstration of power-gating. . . . . 35

Figure 20 Power-gating-efficiency learner schematic. . . . . 35

Figure 21 Power-gating-efficiency learner operation: (a) break-even time tracking and (b) leakage-energy savings $\int \Delta P_{leak} dt - E_{ov}$ tracking. . . . . 37

Figure 22 A demonstration of self-adaptive power-gating. . . . . 38

Figure 23 Test-chip photo-shot and experimental setup. . . . . 39

Figure 24 Measurement results with idle signals of varying $T_{OFF}$ and comparison of power in various modes: (a) LVT design and (b) RVT design. . . . . 40

Figure 25 A Measurement results on break-even point tracking at varying on-chip temperature. . . . . 41

Figure 26 Self-adaptive power-gating across power-density ($T_{OFF} = 2\mu s$), LVT design. . . . . 42

Figure 27 Self-adaptive power-gating (SAPG) for varying activity patterns: (a) measured waveform showing SAPG at randomly varying activity, (b) measured waveform showing the better adaptation at smaller learning cycle, and (c) measured total (leakage + overhead + leaner) power at various learning cycle for different patterns. . . . . 43

Figure 28 PGE Learner's probability to power-gate under dynamic noise. . . . . 44

Figure 29 Implications of PGE learner-inaccuracy: (a) power-penalty in SAPG due to BE inaccuracy, and (b) power-penalty due to dynamic-noise across power densities (temperatures). . . . . 45

Figure 30 Bias-dependences for n-MOSFET in 90nm CMOS: (a) $g_m/I_{DS}$ and (b) $f_T$. [$V_{th}$ = 0.4V] . . . . . 49

Figure 31 Si-Ge heterojunction TFET demonstration: (a) fabricated structure in [1] and (b) equivalent TCAD structure. . . . . 51

Figure 32 III-V heterojunction TFET demonstration: (a) fabricated structure in [2] and (b) equivalent TCAD structure. . . . . 51

Figure 33 Calibration of TCAD simulation parameters against hardware: (a) Si-Ge TFET and (b) III-V TFET. . . . . 53

Figure 34 Comparison of characteristics among Si-Ge TFET, III-V-TFET, and FinFET: (a) $I_{DS}$-$V_{GS}$ and (b) $g_m$-$V_{GS}$. . . . . 55

Figure 35 Comparison of $I_{DS}$-$V_{DS}$ characteristics among Si-Ge TFET, III-V-TFET, and FinFET. . . . . 56

Figure 36 Comparison of capacitance among Si-Ge TFET, III-V-TFET, and FinFET: (a) $C_{GS}$-$V_{GS}$ ($V_{DS}$ = 0.1 V), (b) $C_{GS}$-$V_{GS}$ ($V_{DS}$ = 1 V), (c) $C_{GD}$-$V_{GS}$ ($V_{DS}$ = 0.1 V), and (d) $C_{GD}$-$V_{GS}$ ($V_{DS}$ = 1 V) . . . . . 57

Figure 37 Simulation methodology for TFET-based circuit designs. . . . . 58

Figure 38 (a) Schematic of cross-coupled operational-transconductance-amplifier (OTA). (b) $I_{OUT}$-$V_{IN}$ of the OTA at varying design factor K. . . . . 59

Figure 39 Comparison of OTA-characteristics among SiGe-TFET, III-V-TFET, and FinFET across operating power: (a) transconductnace per power (GM/$P_{OTA}$), (b) output resistance ($R_{OUT}$), and (c) output capacitance ($C_{OUT}$). . . . . 61

Figure 40 CNN synapse power comparison among SiGe TFET, IIIV TFET, and FinFET-based implementations at the varying CNN cell time-constant. . 66

Figure 41 (a) Schematic of a seven transistor OPAMP used in the integrator. (b) For a closed-loop gain of 25, the integrator power at varying R×C time-constant. . . . . 67

Figure 42 Throughput efficiency of SiGe-TFET, IIIV-TFET, and FinFET-based CNN at varying operating power under fixed array approach for (a) edge-detection, and (b) noise-filtering. . . . . 69

Figure 43 Throughput efficiency of SiGe-TFET, IIIV-TFET, and FinFET-based CNN at varying operating power under energy-optimal approach for edge-detection. . . . . 70

Figure 44 (a) Recall probability at varying degree of quantization and for varying NR CNN-AM. (b) Memory capacity and input noise tolerance (in HD bits), at varying NR. Results are for 11×11 CNN-AM. . . . . 73

Figure 45 Cohesive simulation methodology, integrating TCAD, SPICE, and functional simulations to extract CNN-AM characteristics at different technologies, TFET and FinFET. . . . . 74

Figure 46 Transient evolution of various cell state voltages, and the network settling time. . . . . 75

Figure 47 Circuit scheme to locally store and implement quantized synapse weights. 75

Figure 48 (a) Distribution of synapse weights. (b) Recall speed and throughput-efficiency for TFET- and FinFET-CNN-AM across CNN power. . . . . 76

Figure 49 The synapse power distribution and comparison between TFET -and FinFET-CNN-AM for (a) low and (b) high performance application. . . . 77

Figure 50 The OTA transconductance and net CNN cell transconductance across NR for iso-powered CNN designs. . . . . 78

Figure 51 Throughput efficiency (TE) of TFET- and FinFET-CNN-AM at varying NR. . . . . 79

Figure 52 For maximum NR operation between TFET and FinFET CNN-AM at varying power: (a) input pattern noise tolerance, and (b) memory capacity. 79

Figure 53 Probability to recall across NR architectures with increasing variability in: (a) GM and (b) VO (normalized by saturation limit of $f_{sat}$). . . . . 80

Figure 54 Saturating HD noise tolerance and throughput efficiency at increasing number of synaptic interconnections. . . . . 81

Figure 55 Computing elements and architecture of a distance-based neuromorphic-associative-processing. . . . . 85

Figure 56 Requirements of an energy-efficient associative-processing, and contributions in this chapter. . . . . 85

Figure 57 TCAD calibration: (a) a fabricated $In_{0.65}Ga_{0.35}As/GaAs_{0.4}Sb_{0.6}$ n-HTFET in [2], and (b) calibration to the measured data. . . . . 86

Figure 58 Gate/Source-overlapped HTFET (SO-HTFET). . . . . 86

Figure 59 Energy-band diagram and hole-generation profile in SO-HTFET. . . . . 87

Figure 60 Gaussian $I_{DS}$-$V_{GS}$ in SO-HTFET: (a) comparison of $I_{DS}$-$V_{GS}$ of SO-HTFET and a typical HTFET, and (b) comparison of FG-SO-HTFET $I_{DS}$-$V_{GS}$ with the standard-Gaussian characteristics. . . . . 88

Figure 61 Controlling Gaussian-$I_{DS}$-$V_{GS}$ of SO-HTFET: (a) at varying fin-width ($W_{FIN}$), (b) at varying gate/source-overlap length ($S_{OV}$), and (c) at varying channel/source material. . . . . 89

Figure 62 Programming Gaussian-$I_{DS}$-$V_{GS}$ of FG-SO-HTFET: (a) FG-SO-HTFET schematic, and (b) charge-injection to the floating-gate at sufficiently high gate-programming-voltage. . . . . 90

Figure 63 $I_{DS}$-$V_{GS}$ characteristics of FG-SO-HTFET: (a) $V_{peak}$ programming by injecting charge to the floating-gate, and (b) $\Delta V_{peak}$ at varying programming-period ($T_{prog}$) and gate-programming-voltage ($V_{prog}$) [$V_{DS}$ = 0]. . . . . 91

Figure 64 Design of the charge-trapping stack in FG-SO-HTFET: (a) programming-voltages and charge-retention at varying floating-gate thickness, (b) non-uniform charge-trapping in nitride layer-based SO-HTFET, and (c) modulation of both $V_{peak}$ and $I_{peak}$ in a nitride-SO-HTFET. . . . . 92

Figure 65 SO-HTFET-based associative-processing: (a) architecture of SO-HTFET associative-processing array, and (b) column-current. . . . . 93

Figure 66 SO-HTFET-AP peripheral: (a) HTFET winner-take-all, and (b) transients simulations. . . . . 94

Figure 67 On-line training in SO-HTFET-AP: (a) by column programming pulses and test-pattern $V_{test}$ at the array gates, (b) $V_{peak}$ modulation at varying program voltage ($V_{prog}$) [$T_{prog}$ = 100ns], (c) transients for initial $V_{peak}$ = 0 & 1V and $V_{test}$ = 1 & 0V, respectively, and (d) transients showing $V_{peak}$ to follow $V_{test}$ average over training steps. . . . . 96

Figure 68 SO-HTFET-AP-based face-recognition: (a) the simulation methodology, and (b) considered false-positive and false-negative errors. . . . . 98

Figure 69 Accuracy in SO-HTFET-AP-based face-recognition: (a) considered process imperfections in SO-HTFET and imprecision in WTA, (b) accuracy at varying resolution of AP, (c) accuracy in InAs/$GaAs_{0.1}Sb_{0.9}$-SO-HTFET and $In_{0.65}Ga_{0.35}As$/$GaAs_{0.4}Sb_{0.6}$-SO-HTFET-based AP across $V_{test}$ range, and (d) accuracy comparison between Gaussian distance-based AP (as in SO-HTFET-based AP) and L1/L2 norm-based AP (as in the conventional designs [3]). . . . . 99

Figure 70 Reinforced learning in SO-HTFET-AP-based face-recognition: (a) stored pattern plasticity demonstration, (b) improving prediction success with reinforced learning with test iterations, and (c) recognition-accuracy at varying learning-weight ($\alpha$). . . . . 101

Figure 71 Reinforced learning under floating-gate charge leakage: (a) floating-gate charge leakage resulting in stored pattern decay and (b) cumulative recognition accuracy with and without reinforced-learning in a thin-gate SO-HTFET-AP. . . . . 102

Figure 72 Existing AP approaches: (a) digital ASIC, (b) CMOS-analog, (c) floating-gate CMOS-analog, and (d) $VO_2$ oscillators-based. . . . . 103

# CHAPTER 1

# INTRODUCTION

## 1.1 A Paradigm-Shift in the Contemporary Computing: Added Emphasis to Energy-Efficiency

Projected near future growth rate for various computing-platforms is presented in Figure 1. The traditional high-performance computing-platforms such as servers and desktops are only showing a limited growth; meanwhile, a remarkable growth is anticipated in the low-power computing-platforms: smart-phones, internet-of-things (IoTs), and wearables.

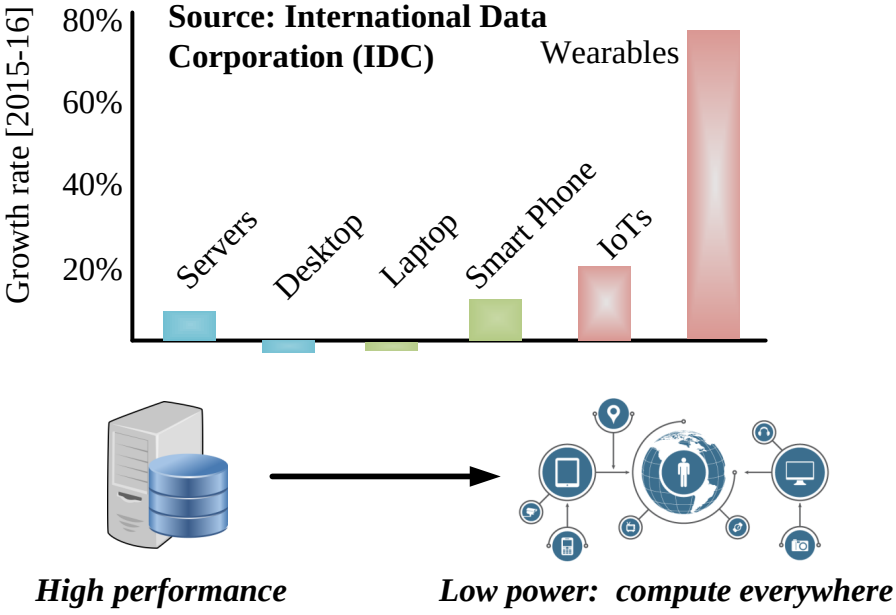



**Figure 1: Near future growth-rate of various computing-platforms.**

The remarkable growth in the low-power computing-platforms is a result of their phenomenal application prospects. By their virtue of 'low in power-consumption', the low-power computing platforms can often be powered by a battery and be placed non-intrusively

in any utility of daily use. Such distributed and mobile deployment of the computing platforms has dramatically improved the quality of our lives. For instance, wearable health-instruments are transforming health-care [4]; and seamlessly engaged home-appliances are making a home smarter [5]. As a result, the number of mobile computing devices has surged, and has now even surpassed the human population on earth. However, as the prospects of low-power computing grow, a variety of newer applications are demanding not only a mere low-power operation but also a higher-performance one. For example, Amazon is developing battery-powered drones to deliver packages which will require to process complex navigational schemes in real-time. Another emerging computing-platform, internet-of-things, is interconnecting several computing elements, and is escalating the scale and complexity of data-processing in the computing devices. Moreover, the electronic-computing has evolved beyond its conventional practices in arithmetic and storage; and the computing is now deployed in the diverse computational task such as associative-processing and high-dimensional information-processing. Therefore, to sustain the growth of such applications and their startling impact on our lives, the energy-efficiency (i.e., a higher performance per power) of the computing-platforms has become the foremost design-constraint.

## 1.2 Energy-efficient Computing Approach of the Dissertation

The approach to optimize computing energy-efficiency pursued in this dissertation is two-fold:

### 1.2.1 Explore the Role of Emerging Transistor Technologies

At first, the dissertation focuses on the most elementary component of any electronic design, i.e., a transistor, to enhance the energy-efficiency of computing. One of the most fundamental limitation of the conventional metal-oxide-semiconductor field-effect-transistor (MOSFET) is a limited switching-slope (SS). A limited SS in MOSFET is due to its thermionic-injection-based charge-conduction. A limited SS of MOSFET leads to limited

energy-efficiency of MOSFET-based digital and analog designs. The dissertation investigates tunneling field-effect-transistors (Tunnel-FET or TFET) for energy-efficient computing. A TFET, unlike MOSFET, conducts by band-to-band-tunneling (BTBT)-based conduction. A BTBT conduction in TFET achieves much steeper SS than MOSFET. Specifically, the role of steeper SS of TFET in improving energy-efficiency of analog-designs is shown.

### 1.2.2 Explore the Role of Application-Specific and Intelligent Hardware

Secondly, the dissertation investigates the role of application-specific and intelligent computing-platforms for emerging computing demands. A neuromorphic computing is an emerging computing-paradigm inspired by the neuron-based dynamics in a biological brain. Neuromorphic-computing architectures such as cellular neural network (CNN) have shown dramatic improvements in energy-efficiency from the conventional architectures, particularly in the non-conventional tasks, such as image processing, associative memories, and classification []. Intriguingly, unique and rich characteristics of TFET also holds the promise to significantly reduce complexity and increase energy-efficiency of various non-conventional computing architectures. Therefore, the dissertation explores the potential of a blend of TFET and non-conventional computing architectures. Specifically, TFET-based cellular neural network and distance-computing-based associative-processing are investigated.

## 1.3 Contributions of the Dissertation

The dissertation follows emerging technologies and non-conventional computing architectures for energy-efficient computing, and makes the following contributions:

- **MOSFET-based single neuron learning and application to minimal wasted-power in digital computing**: MOSFET-based realization of a single neuron learning circuit

is shown. The neuron-circuit optimally configures a digital system to the power-gated mode (i.e., the low-leakage mode) or the non power-gated model (i.e., the typical mode) to ensure minimal wasted-power dissipation. The neuron circuit considers mode-transition energy trade-offs against leakage-savings, and learns and adapts in runtime against varying input-activity, process, and temperature conditions. A test-chip in 130nm CMOS demonstrates single neuron-based minimal power-operation. As compared to the existing power-gating controllers, the presented single neuron-based controller incurs the minimum area and power overheads while featuring history-based and process/temperature-based adaptations.

- **TFET-based multi-neuron cellular neural network and application to energy-efficient image-processing and associative-memory**: The application of multi-neuron-computing based CNN architecture for image-processing and associative-memory is shown. The limitations of MOSFET-based CNN are illustrated, and the role of TFET to improve energy-efficiency of CNN is examined. Various optimal design approaches for TFET-CNN-based image-processing and associative-memory are demonstrated. At the same operating-power TFET-CNN-based image processing achieves more than 10× improvement in edge-detection and noise-filtering throughput-efficiency than the FinFET-CNN-based design. At the same operating-power TFET-CNN-based associative memory achieves 10× improvement in memory-capacity and 3× improvement in noise-tolerance than the FinFET-CNN-based design.

- **Non-conventional TFET design for neuromorphic-computing and application to face-recognition**: A non-conventional gate/source-overlapped heterojunction Tunnel-FET (SO-HTFET) with Gaussian-$I_{DS}-V_{GS}$ characteristics is shown. The SO-HTFET designs a single-transistor distance-computing-cell (DCC) for associative-processing (AP) with reinforced-learning. The application of SO-HTFET-based AP to face-recognition demonstrates a higher-accuracy, 250× lower-power, and 100× higher

DCC-density than a digital-CMOS-based Boolean-AP.

## 1.4 Organization of the Dissertation

The organization of this disseration is as following. Various literature works illustrating fundamental limitations of MOSFET, desing and characteristics of TFET, and neuromorphic-computing are reviewed in Chapter 2. In Chapter 3, emerging process variability and reliability challenges in non-conventional MOSFET design and integration are investigated. In Chapter 4, a single-neuron based power-gating controller is shown to alleviate leakage-power limitations of MOSFET. Chapter 5 discusses the role of TFET in low-power analog-computing. Specifically, TFET-based ultralow power operational transconductance amplifier (OTA) design is developed. In Chapter 6, TFET-CNN-based energy-efficient image-processing is discussed. In Chapter 7, TFET-CNN-based energy-efficient associative-memory is discussed. Chapter 8 shows non-conventional TFET design (i.e., SO-HTFET) and a single-transistor DCC based on SO-HTFET. SO-HTFET-DCC based AP and face-recognition is also shown. Chapter 9 summarizes and concludes this dissertation.

# CHAPTER 2

# LITERATURE SURVEY

A survey on various computing-platforms emphasizing a paradigm shift in the contemporary computing demands is first shown in this chapter. Afterwards, the need for neuromorphic-computing to address such emerging-computing demands is examined. The advantage of emerging-technologies in energy-efficient neuromorphic-computing platforms is further illustrated. Specifically, this dissertation considers the role of emerging tunneling-field-effect-transistors (Tunnel-FETs or TFETs) for neuromorphic-computing. Therefore, the key-attributes and recent-developments in TFET-technology are subsequently presented.

## 2.1 Switching-Slope-induced Energy-Efficiency Limitations in MOSFET-based Computing

A limited switching-slope (SS) is one of the most fundamental deficiencies of MOSFET leading to limited energy-efficiency in MOSFET-based design. MOSFET conducts by thermionic carrier-injection, and SS of a MOSFET in its subthreshold or weak-inversion region can be expressed as [6]

$$SS = \frac{dV_{GS}}{dlog_{10}I_{DS}} = \frac{dV_{GS}}{d\psi_s}\frac{d\psi_s}{dlog_{10}I_{DS}} = \left(1 + \frac{C_{dep}}{C_{ox}}\right) \times ln10\frac{kT}{q}. \quad (1)$$

Here, $V_{GS}$ is the gate-to-source voltage; $I_{DS}$ is the drain-to-source current; $\psi_s$ is the surface potential; $C_{dep}$ is the depletion capacitance; and $C_{ox}$ is the oxide capacitance in MOSFET. By optimizing MOSFET design and operation to minimize $C_{dep}/C_{ox}$, SS in MOSFET can be minimized. However SS is still limited by the thermal voltage $V_T = kT/q$. At room temperature, SS is greater than 60mV/decade. In realistic designs, SS is limited by the other factors such as the drain fringing fields to channel where SS ~70-80mV/decade is achieved.

### 2.1.1 Limitations in MOSFET-based Digital-Computing

A limited SS of MOSFET, in turn, limits the minimum switching energy of MOSFET-based digital-designs. The total operational energy ($E_{Total}$) of a digital-design can be expressed as [6]

$$E_{Total} = E_{dynamic} + E_{leakage} = L_d \times \left(\alpha \cdot C \cdot VDD^2 + I_{OFF} \cdot VDD \cdot \tau_{delay}\right)$$
$$\approx L_d \times C \times VDD^2\left(\alpha + 10^{-VDD/SS}\right) \quad (2)$$

Here, $L_d$ is the logic depth, $C$ is the total switching capacitance, $\alpha$ is the switching activity, and $\tau_{delay}$ is the switching time in the logic block. Therefore, with a limited SS in MOSFET, the minimum switching energy in MOSFET-based digital designs and their energy-efficiency is limited.

### 2.1.2 Limitations in MOSFET-based Analog-Computing

A limited SS also limits the energy-efficiency of MOSFET-based analog designs. Various works operate analog designs under subthreshold bias conditions to minimize power dissipation and to enhance transconductance per bias current ($g_m/I_{DS}$) for better energy efficiency. In the subthreshold region, $g_m/I_{DS}$ in MOSFET is related to SS as [7]

$$\frac{g_m}{I_{DS}} = \frac{ln(10)}{SS} \leq \frac{q}{kT} = V_T \quad (3)$$

Therefore, with a limited SS, $g_m/I_{DS}$ of MOSFET is less than $V_T$ and the energy efficiency of MOSFET based analog designs is also limited.

### 2.1.3 Emerging Transistor Technologies with Steeper Switching-Slope

A number of emerging-transistors with sub-thermal SS are being investigated to overcome the energy-efficiency limitation of MOSFET. TFET devices conducting on the principle of band-to-band-tunneling have shown SS as low as 36mV/decade [8]. However, ambipolarity and higher miller capacitance in TFET is a challenge [9]. A negative gate capacitance was explored by embedding ferroelectric layer in the gate insulator. Through a negative gate

capacitance, ferroelectric FET can evade the thermal limit of SS; however, drawbacks are hysteresis and a larger switching time. Molecular transistors with steeper SS were investigated where the gate voltage either modulates the coupling of the channel molecules to the source/drain contacts or varies the transmission properties of the channel molecules itself [10]. However, SS of a molecular FET is significantly inferior when considering metallic contacts [11]. Impact ionization MOS (IMOS) conducting through impact ionization of carriers have been demonstrated with a very small SS. However, IMOS suffers from a hot carrier injection induced reliability issues [12]. Feedback FET (FBFET) utilizes positive feedback induced by the charge injection to the dielectric spacer; however, a strong hysteresis in the transistor is a challenge [13]. NEMFET uses a mechanical force to physically disconnect gate from the dielectric and achieves a very small SS; though, the mechanical movement of gate results in excessively larger switching times [14].

Among the emerging transistors with steeper SS, a variety of works have shown a greater interest towards TFET for the next generation replacement of MOSFET [15, 6]. CMOS compatible fabrication process steps in TFET are promising for a large scale integration [16]. TFET devices based on heterojunction have demonstrated MOSFET-like $I_{ON}$ while achieving a very low $I_{OFF}$ [17]. Ambipolarity of TFET can be suppressed by employing drain underlap [18], and higher miller cap can be reduced by channel material of lower density of mass [9]. TFET also exhibits suppressed temperature dependence [19]. Since the current conduction in these transistors is controlled by the source/channel interface, TFETs are relatively immune to short channel effects, are more scalable, and show a near perfect saturation [7]. Hence, arguing TFETs to be the next generation MOSFET replacement, this dissertation investigates background and related work in TFET in the subsequent section.

## 2.2 Tunneling-Field-Effect Transistors (Tunnel-FET or TFET)

In the quest of achieving higher energy-efficiency, Tunnel FET (TFET) as an alternative to the conventional transistor, MOSFET, is being explored. In contrary to a thermally-limited

switching-slope (SS) in MOSFET, TFET can achieve subthermal SS. The steeper SS in TFET achieves a lower supply voltage (VDD) operation than MOSFET without sacrificing $I_{on}/I_{off}$ ratio. VDD scaling of TFET based digital-circuits leads to a quadratic reduction in the dynamic power. TFET can also achieve very low-off current as compared to a MOSFET thanks to its diode like built-in barrier. Thus, the standby power of TFET-based digital circuits will also be very low. The above attributes of TFET have made it the leading emerging-technology of interest for low-power and energy-efficient computations. The key aspects of TFET are reviewed here.

### 2.2.1 Charge-Conduction in TFET

In Figure 2a, a TFET, similar to a MOSFET, is comprised of source, channel, and drain regions; however, unlike MOSFET, the doping of source and drain regions is different from the each other. In an n-TFET, the source is doped as $P^+$ type and the drain doping is $N^+$. While in a p-TFET, the source is $N^+$ doped and the drain is $P^+$ doped. The channel in TFET is intrinsic. In an n-TFET, current conducts by electron-tunneling from the valence-band in source to the conduction-band in drain [Figure 2b]; while in a p-TFET, the holes tunnel from the conduction-band in source to the valence-band in drain. This band-to-band-tunneling (BTBT) depends on the barrier-height and width as shown in Figure 2b. The barrier-height is determined by the material bandgap at the source/channel junction, whereas the barrier-width can be controlled by the gate-electrode potential. As shown in Figure 2b, in the off-state ($V_{GS}$ = 0V), the barrier-width is significantly large, and thereby, BTBT probability is very low, and the off-current in TFET can be significantly low. However, as the gate-voltage increases, a significant band-bending occurs at the source/channel junction and the barrier width decreases. Hence, with the decreasing barrier-width at the high gate-voltage, a significant charge injection from source to channel and current-conduction to drain occurs in TFET.

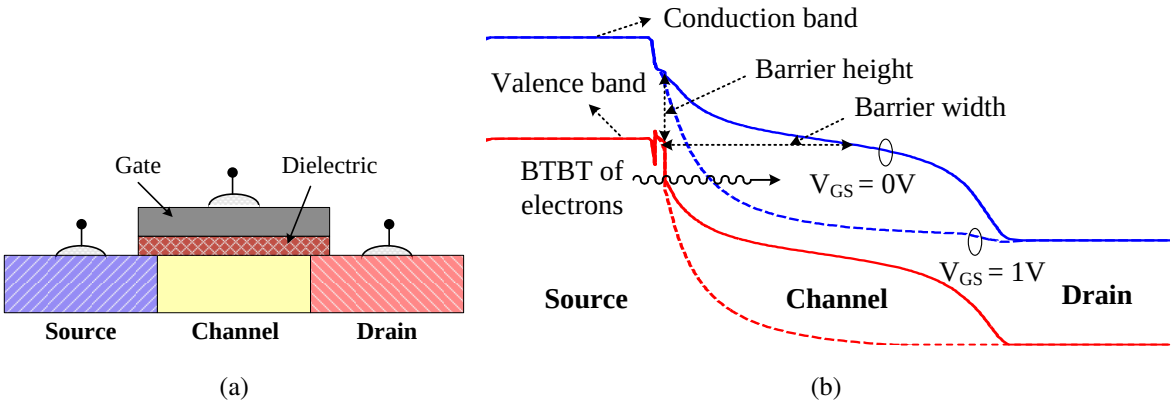



**Figure 2: (a) TFET schematic. (b) Energy-band diagram and band-to-band-tunneling in TFET.**

Because of a high bandgap in silicon, the barrier-height and barrier-width of BTBT in a silicon-channel TFET is high, and $I_{ON}$ is very low. The efforts to improve on-current of TFET are primarily two-fold. A group of researchers investigate a variety of techniques to improve the on-current in silicon-channel TFET. Meanwhile, another group of researchers are exploring TFETs with beyond silicon low-bandgap channel materials. These various approaches are summarized in Figure 3. Although, non-silicon channel TFETs are promising to significantly improve on-current of TFET, a higher off-current is also incurred. The key-contributions in both of these approaches are reviewed subsequently.

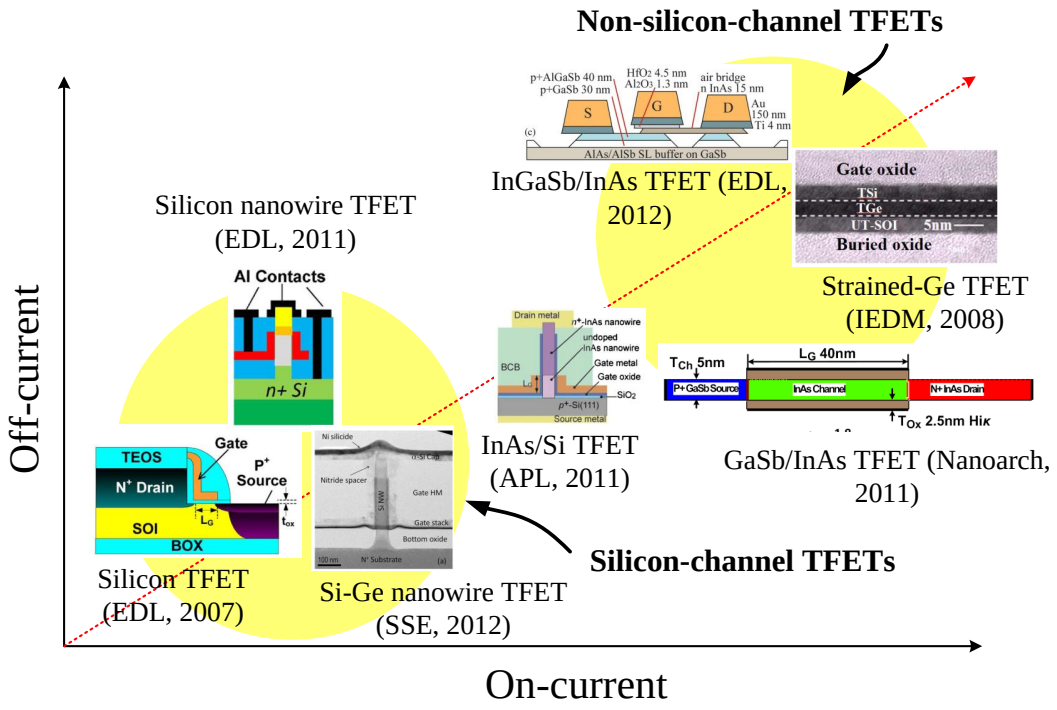



**Figure 3: Silicon-channel and non-silicon-channel approaches to improve on-current of TFET.**

### 2.2.2 Silicon-channel TFET

Silicon TFETs are attractive because of a compatible fabrication and co-integration with CMOS, however, a low on-current in silicon-TFET is a challenge. Boucart et al. suggested using stress to lower the bandgap at the source/channel junction leading to $I_{ON}$ enhancement [20]. A thin Si/Ge layer was embedded at the source/channel junction, where due to the lower-bandgap of Si/Ge layer than Silicon, $I_{ON}$ in TFET increases [21]. Bilayer silicon TFET was investigated, where the p-i-n band bending of TFET is directly aligned with the gate electric-field, and the improved gate-electrostatics increases $I_{ON}$ in TFET [22]. Source-silicidation was investigated to increase the junction steepness at the source/channel junction to improve $I_{ON}$ in TFET [23], while abrupt doping profile in axially doped silicon-TFET was achieved [24]. Halo-doping at the source/channel junction [25] was utilized to improve $I_{ON}$ in silicon TFET. Raised Ge-source silicon-channel TFET with the higher $I_{ON}$ were shown [26]. A silicon-channel vertical nanowire TFET with gate-all-around and

embedded Si/Ge layer at the source/channel junction was demonstrated to improve $I_{ON}$ and SS in [27]. Among the various techniques, a vertical TFET with the embedded Si/Ge layer is the most attractive due to a simplicity in fabrication steps, and a significant enhancement in on-current.

### 2.2.3 Alternate Channel-Materials for TFET

Several low-bandgap channel materials have been investigated to overcome the on-current deficiencies of silicon-TFET. Meyer at al. reported much higher $I_{ON}$ with GeOI TFET devices [28] than in silicon-TFET. Bilayer graphene-based TFET demonstrated ultralow SS ($\leq$ 20mV/dec) [29]. TFET-based on InAs/GaSb heterojunction demonstrated MOSFET like $I_{ON}$ and suitability for hybrid-core integration with CMOS [30]. InAs-Si nanowire heterojunction TFETs were shown with much reduced temperature dependence. $I_{ON}$ of InAs TFETs was improved with strain engineering [31]. Computational study of carbon nanotube p-i-n TFETs demonstrated their potential for high speed ultralow power sub 0.4V VDD logic operation [32].

## 2.3 Non-Boolean Computing Architecture

Conventional digital computing is Boolean in nature where inputs & outputs are binary coded. A non-Boolean computing, however, doesn't necessitate a binary coded inputs/outputs and also processes a continuous or discreet inputs/outputs. For a variety of computing-tasks, the non-Boolean computing can be much more energy-efficient than a Boolean computing. A salient case of non-Boolean computing is information processing in a biological brain. Non-Boolean computing architectures can also be highly process variability resilient, and are much more amenable to feature-size scaling. This dissertation primarily explores two classes of non-Boolean computing architectures, a cellular neural network and a grid-based associative processing architecture. Brief review of these architectures is given below.

### 2.3.1 Cellular Neural Network (CNN)

Cellular neural network (CNN) is a computational architecture inspired by the neuron and synapse-based dynamics of a biological brain. A cellular neural network (CNN) platform is composed of a set of neurons organized as a 2D-array where each neuron is inter-connected to its local neighbors using synapses [33] [Figure 4]. This local connectivity makes CNN an attractive hardware platform, particularly in advanced nanometer nodes where interconnect scaling is challenging [34]. Each cell ($C_{ij}$) in CNN consists of three nodes: input ($u_{ij}$), state ($x_{ij}$), and output ($y_{ij}$). Underlying dynamics of the CNN cell $C_{ij}$ is given by [35]

$$C\frac{dV(x_{ij})}{dt} = -\frac{V(x_{ij})}{R} + \sum_{kl\in S_{ij}} A_{kl,ij} f_{act}(V(x_{ij})) + \sum_{kl\in S_{ij}} B_{kl,ij} V(u_{kl}) + I_{ij}, \tag{4a}$$

$$y_{ij} = f_{act}(x_{ij}), \tag{4b}$$

where $S_{ij}$ is the set of neighboring cells directly connected to the cell $C_{ij}$. $f_{act}$ is the activation function with sigmoid characteristics and saturation limits, $V_{TP}$ and $V_{TN}$. In CNN dynamics, at equilibrium the output saturates to the limits of $f_{act}$, $V_{TP}$ or $V_{TN}$ [35]. The parameters $A_{kl,ij}$ and $B_{kl,ij}$ represent the feedback and feed-forward templates, and the parameter $I_{ij}$ represents the bias term. These template weights are programmed to realize various functionality of CNN. Figure 4a demonstrated a neighborhood radius (NR) of one implementation, while a higher NR implementation can be similarly done by interconnecting more cells together. CNN neuron interconnection for the NR of two are shown in Figure 4b.

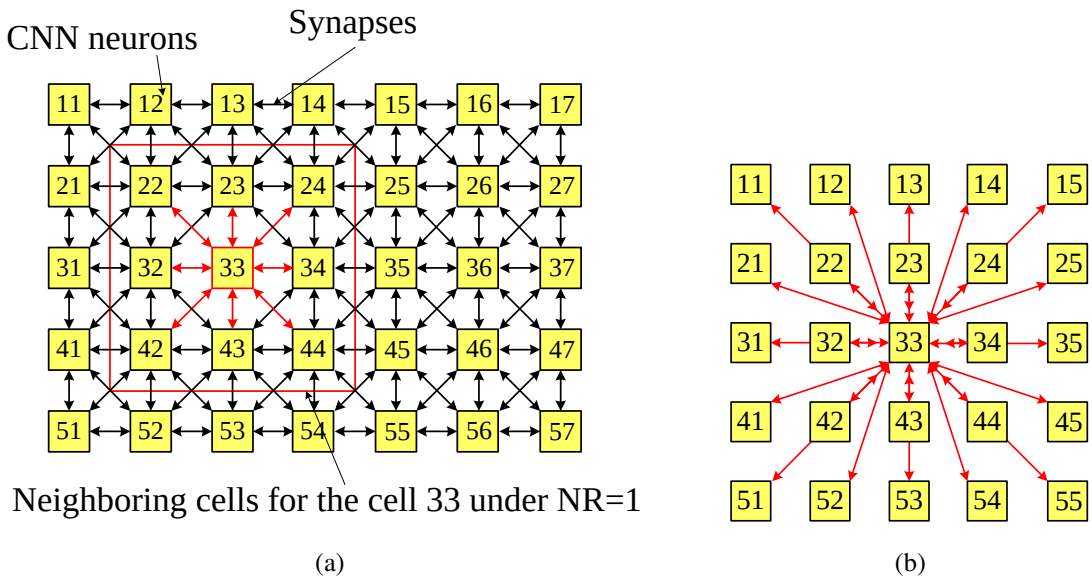



**Figure 4: (a) Cellular neural network (CNN) schematic. (b) CNN neuron interconnections for the neighborhood-radius of two.**

An analog implementation of CNN, based on the design in [36], is presented in Figure 5. The neuron of CNN is implemented using an operation amplifier (op-amp)-based integrator including resistor & capacitor (RC) elements. A current-source generates bias, $I_{ij}$, of the neuron. A unity-gain op-amp with the saturated output was shown to implement sigmoid activation-function ($f_{act}$) in [36]. A synapse to enable inter-neuron interaction is implemented with operational-transconductance-amplifier (OTA). The OTA-transconductance matches $A_{kl,ij}$ and $B_{kl,ij}$ of the neuron.

CNN is a versatile computing platform, and it can be used in a number of computing tasks by appropriately programming synaptic weights, $A_{kl,ij}$ and $B_{kl,ij}$ and bias $I_{ij}$. Application of CNN to image-processing and associative-memory is discussed in the following.

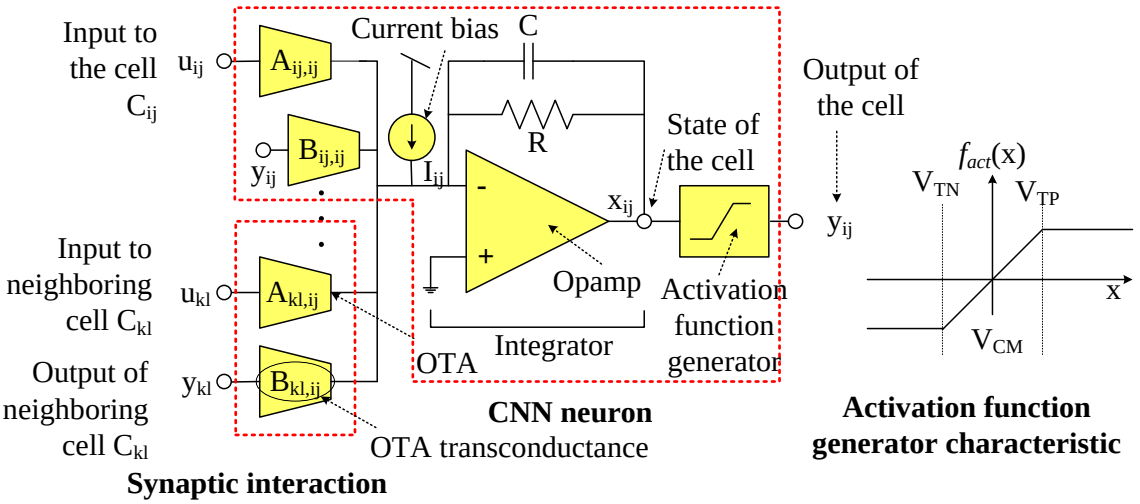



**Figure 5: An analog implementation of cellular neural network.**

### *2.3.1.1 CNN-based Image Processing*

CNN-based image-processing is enabled by appropriately programming CNN-parameters: $A_{kl,ij}$, $B_{kl,ij}$, and $I_{ij}$. The programming parameter templates for various CNN-based image-processing operations were developed in [35]. The parameter-templates are space-invariant, i.e., each CNN-cell is operated with the same template. CNN-based image processing and parameter-templates for edge-detection and noise-filtering is shown in Figure 6. Here, each pixel of the image is mapped on to the CNN-cell input. And, through appropriate CNN-parameters and CNN-dynamics, the processed pixel value evolves at the CNN-cell output. For patterns larger than the CNN array, CNN multiplexes it in a lexicographic fashion (i.e. left to right and then top to bottom). Overlap (= 2×NR pixels) between the successive steps is required to account for the interaction from the neighboring pixels.

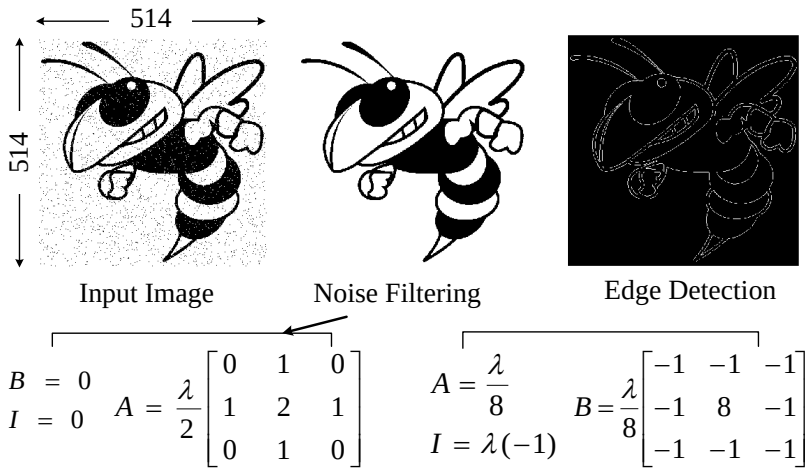



**Figure 6: Cellular neural network-based image-processing: edge-detection and noise-filtering.**

#### *2.3.1.2 CNN-based Associative Memory*

CNN-based AM operation is shown in Figure 7, where the association of letters 'T', 'C' has been established to 'E', 'H', respectively. Various pixels correspond to the CNN cells, and black and white color corresponds to $V_{TP}$ or $V_{TN}$ voltages. When the input of the CNN cells are excited with a distorted pattern of 'C', the output of the cells accurately evolve to the respective pattern 'H' at equilibrium. The design of CNN-AM involves the algorithmic synthesis of CNN-parameters: $A_{kl,ij}$, $B_{kl,ij}$, and $I_{ij}$. Various synthesis algorithms for AM on CNN have been shown, such as, pseudo-inverse based method [37], singular value decomposition based method [38], and Hebbian learning based method [39]. Different methods vary in terms of computation cost and robustness of synthesis [40]. Based on the Hebbian learning rule, method from [39] was shown to be sufficiently robust and with moderate computation cost. A hetero-associative implementation of this method is shown in Figure 8. Notably, while CNN parameters $A_{kl,ij}$, $B_{kl,ij}$, and $I_{ij}$ are space-invariant in CNN-based image-processing, these parameters vary from the cell-to-cell in CNN-based AM.

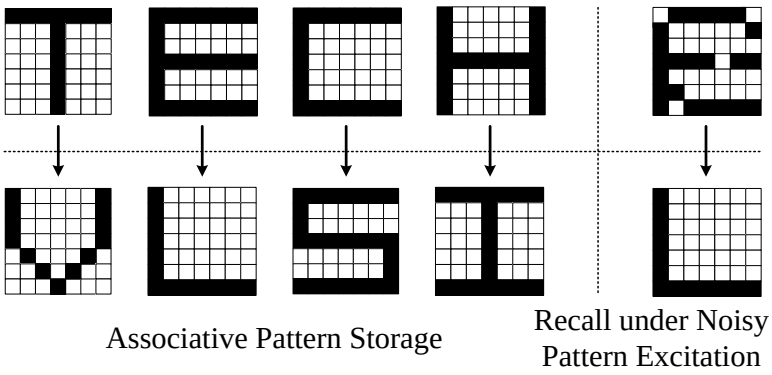



**Figure 7: Cellular neural network-based associative-memory operation.**

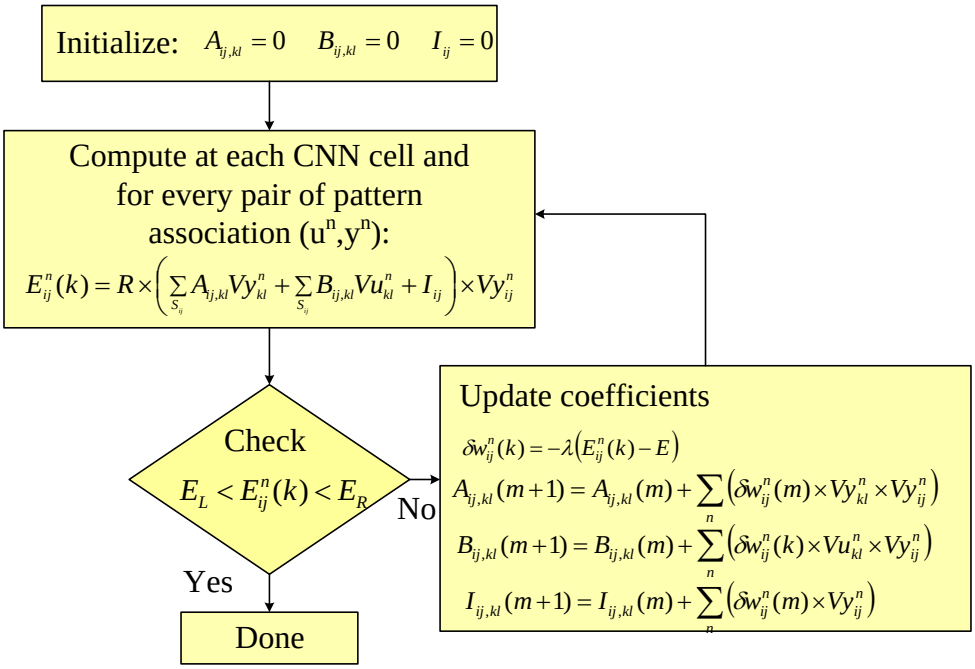



**Figure 8: A Hebbian learning-based hetero-associative recall algorithm.**

### 2.3.2 Grid-based Non-Boolean Associative Computing with Emerging Devices

A CNN-based robust associative computing requires a much higher inter-cell interconnections leading to much greater implementation complexity. Beyond CMOS unique characteristics of various emerging devices have been applied to simplify cell design and realize a highly interconnected network with reduced complexity. Particularly, regular grid based architectures have been developed where all the computing cells within a column (or, in

the whole array) are interconnected. Spintronics of magnetic domain-wall-memories was exploited in a compact, energy-efficient associative computing [41]. Metal-to-insulator transition in $VO_2$ was exploited in a coupled relaxation oscillatory network for associative computing [3]. Negative gate-transconductance of single-electron-transistor (SET) was exploited in an associative computing platform [42]. As well, CMOS and emerging-technologies-based compact and regular peripherals have been designed for the operation of a grid-based non-Boolean computing. Winner-take-all (WTA) circuit was demonstrated in [43] to detect the highest current magnitude column in an array. Another WTA design combining the unique features of spin magnetic-tunnel-junction (MTJ) with the MOSFET-based circuitry was proposed in for higher precision and variability tolerance [41]. This dissertation, similar to the prior works, follows a regular grid-based neuromorphic computing platform. However, unlike the above approaches, conventional computing transistors are re-engineered for grid-based neuromorphic computing to facilitate seamless integration between the conventional computing and neuromorphic computing. Specifically, the dissertation introduces a non-conventional TFET design (called as SO-HTFET), and demonstrates a grid-based non-Boolean associative computing with SO-HTFET.

## 2.4 Conclusions

Limitations in MOSFET-based digital and analog computing arise due to a limited switching slope in MOSFET. This dissertation in the subsequent chapters explores the potential of an emerging subthermal swing TFET technology to overcome these energy-efficiency limitations of MOSFET. Furthermore, as the electronic computing evolves to unconventional tasks such as associative memory, there is a critical need to find more suitable computing architectures for these unconventional tasks. Beyond CMOS characteristics of the emerging technologies when combined with the suitable non-traditional computing architectures can unlock much higher energy-efficiency for these unconventional and increasingly prevalent tasks. In pursuit of such higher energy-efficiency for various emerging applications,

this dissertation explores TFET-based cellular neural networks for image processing and TFET-based non-Boolean associative computing in the subsequent chapters.

# CHAPTER 3

# EMERGING DEVICE CHARACTERISTICS VARIABILITY MECHANISMS IN NON-TRADITIONAL MOSFET DESIGN AND INTEGRATION

Limited switching-slope (SS) and process-variability induced energy-efficiency limitations in MOSFET-based design were discussed in Chapter 2. A variety of device design and integration techniques are being pursued by researchers elsewhere to minimize the power-dissipation and improve the energy-efficiency of MOSFET-based designs. However, various device-to-device variability sources lurk in in these non-traditional device design and integration techniques. This chapter explores these emerging device variability mechanisms in the non-traditional MOSFET design and integration. The contributions discussed in the chapter were published in [44, 45].

## 3.1 Threshold-Voltage Variability in Non-traditional High-$\kappa$ Dielectric MOSFET

A high-$\kappa$ dielectric/metal gate-stack has replaced the conventional $SiO_2$/Polysilicon gate-stack in the new generation MOSFET devices. However, the effective gate workfunction (WF) in $HfO_2$/metal gated transistors is observed to shift from its theoretical value [46]. The anomalous WF shift is attributed to the charged defects, oxygen vacancies (OVs). An OV is a thermodynamic point defect caused by the diffusion of oxygen from $HfO_2$, which leaves behind a doubly charged vacancy defect, $V_O^{++}$ [Figure 9]. The presence of positively charged $V_O^{++}$ alters the gate electrostatics. Furthermore, in nanoscaled transistors, as the count and spatial allocation of OVs varies from the device-to-device, OVs also induce a significant local variability in WF, and hence, in threshold voltage ($V_{th}$).

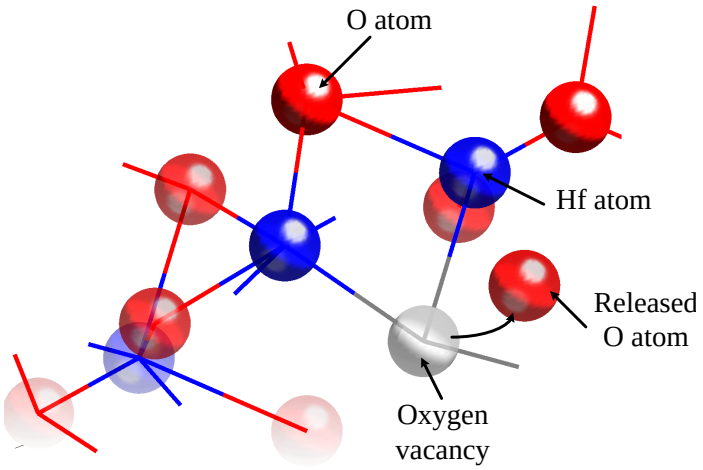



**Figure 9: A demonstration of oxygen-vacancy generation in $HfO_2$.**

### 3.1.1 Oxygen-vacancy Generation Model

The generation of OV can be expressed as

$$O_O^x \leftrightarrow V_O^{++} + 2e + \frac{1}{2}O_2 - \Delta G_1^0. \tag{5}$$

Here, $O_O^x$ is an oxygen atom in the dielectric $HfO_2$, e is the delocalized surplus electron created by OV formation, $O_2$ is the oxygen gas molecule, and $\Delta G_1^0$ is the standard free energy of exchange in OV formation. At equilibrium, according to the law of mass action, the effective OV concentration is established as

$$\frac{[V_O^{++}]}{[O_O^x]}[e]^2 p_{O2}^{1/2} = exp(-\frac{\Delta G_1^0}{kT_{G,form}}). \tag{6}$$

Here, the quantities in the brackets [] represent the corresponding concentration. $T_{G,form}$ is the gate-stack formation temperature. OV-liberated electrons are collected by the adjacent metal electrode. Hence, using Fermi-Dirac statistics [47], the electron concentration can be expressed as

$$[e] = \frac{1}{1 + exp(\frac{E_{OV}(r) - E_{F,m}}{kT_{G,form}})} \approx exp(-\frac{E_{OV}(r) - E_{F,m}}{kT_{G,form}}). \tag{7}$$

Here, $E_{F,m}$ is the metal Fermi energy level. Using Equations 6 & 7, the probability of OV generation will be given as

$$P_{OV}(r) = \frac{[V_O^{++}]}{[O_O^x]} = \frac{1}{p_{O2}^{1/2}} exp(-\frac{\Delta G_1^0}{kT_{G,form}} + 2 \times \frac{E_{OV}(r) - E_{F,m}}{kT_{G,form}}). \quad (8)$$

However, due to its positive charge, an OV affects the potential field and defect energy level, $E_{OV}$, in its proximity, as shown in Figure 10.

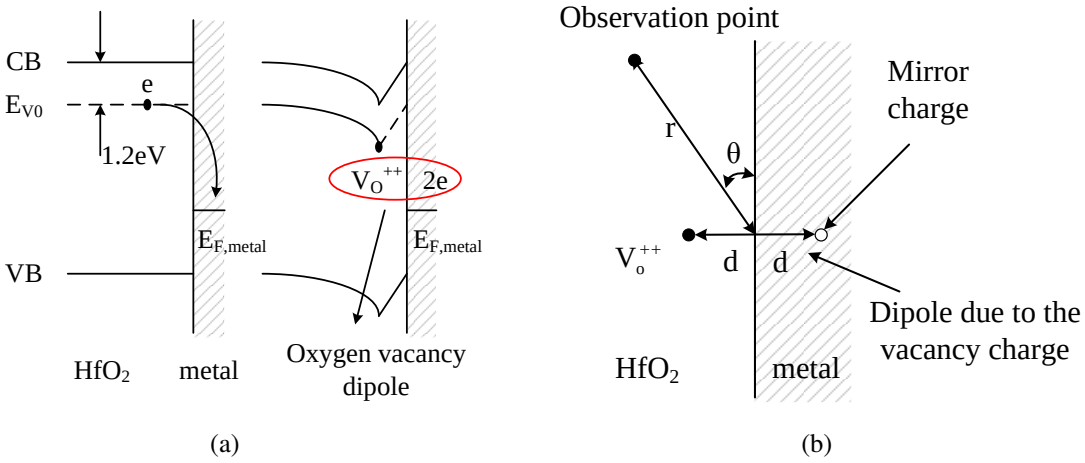



**Figure 10: (a) Surplus electrons fall from the trap energy level into the lower energy metal electrode, and OV dipole forms. (b) OV induced dipole created by the mirror charge.**

With the local variation in $E_{OV}$ relative to the metal Fermi level, $E_{F,m}$, the generation of another OV in proximity is also affected as is evident from Equation 8. The potential field due to the positively charged OV can be analyzed by the method of mirror charges [48], where an opposite and equidistant charge from the metal interface is considered, as shown in Figure 10b. Considering the dipole formed by the OV and its mirror charge, the potential field of an OV dipole can be approximated as

$$\Delta V(r) \approx \frac{q}{4\pi\epsilon} \frac{2d \times sin\theta}{r^2}. \quad (9)$$

The position of the dipole relative to the observation point $r$ is shown in Figure 10b, along

with the notations, $d$, $r$, and $\theta$. Thus, accounting for the potential field of existing OV-induced dipoles ($\Delta V_i(r)$), the probability of OV generation is expressed as

$$P_{OV}(r) = \frac{[V_O^{++}]}{[O_O^x]} = \frac{1}{p_{O2}^{1/2}} \times exp(-\frac{\Delta G_1^0}{kT_{G,form}} + 2 \times \frac{E_{OV}(r) - \Sigma\Delta V_i(r) - E_{F,m}}{kT_{G,form}}). \quad (10)$$

### 3.1.2 Simulation Methodology for Oxygen Vacancy-induced Variability

The OV generation probability expression developed in Equation 10 is used to mimic the random placement of OVs in a $HfO_2$ layer. Films with the dimensions of the $HfO_2$ layer, as in a transistor, are meshed with the node density ~55/nm$^3$. This is equivalent to the oxygen atom density in the $HfO_2$ layer [49]. Each mesh node is considered for possible placement of an OV. To emulate the amorphous nature of $HfO_2$ film, the order in which the mesh nodes are considered is randomized. Using the gate-stack process parameters ($T_{G,form}$, $p_{O2}$, and gate WF), the expression in Equation 10 is evaluated, and an OV is placed based on the probability $P_{OV}$(r). Based on [47], for ultrathin films of $HfO_2$, the value $G_1^0$ = 3 eV is used. As discussed in [47], this value of $G_1^0$ explains the experimental observations for the $HfO_2$/metal gate-stacks in [50]. The energy level $E_{OV}$ is 1.2eV from the $HfO_2$ conduction band, and it corresponds to the donor level of the doubly charged vacancy, $V_O^{++}$. Based on *ab initio* calculations for a-$HfO_2$, the trap charge transition level from the neutral OV to $V_O^{++}$ was computed as ~1.1eV below the $HfO_2$ conduction band [51]. Likewise, based on absorption spectra, the trap energy level of the $HfO_2$ films deposited on silicon was found to be ~1.2eV from the $HfO_2$ conduction band [52]. In agreement with the above work, here I have taken the donor energy level ($E_{OV}$) as 1.2eV below the $HfO_2$ conduction band, since, as discussed subsequently, it also corroborates the OV concentration from the discussed model to the experimentally observed concentration in [53]. The OV distribution/concentration in the successive discussion is obtained for the process parameters: WF = 4.7eV (TiN vacuum value), $p_{O2} = 5\times10^{-8}$ atm, and $T_{G,form}$ = 1300K (the gate-first process) and 750K (the gate-last process).

Each OV is represented by a charge cube (dimension: 0.5 nm), and charge density

equivalent to two electron charges. Using Sentaurus Structure Editor [54], these charged cubes are placed in the $HfO_2$ layer with the corresponding spatial distribution, as shown in Figure 11. Simulations of electrical characteristics were performed using Sentaurus device [54]. Unified mobility models from [55] were used. Scattering models for the mobility degradation at the channel/gate interface due to high-$\kappa$ dielectric were used [56]. Quantum confinement models were used in the channel. A sample distribution of OV-induced surface potential variability at the oxide/channel interface is shown in Figure 12. A significant local variability in the potential, ~0.4-0.65V, is observed. Sharper transitions correspond to the vacancies located closer to the channel. Due to their greater electric field at the oxide/channel interface, a greater gradient in the potential field occurs.

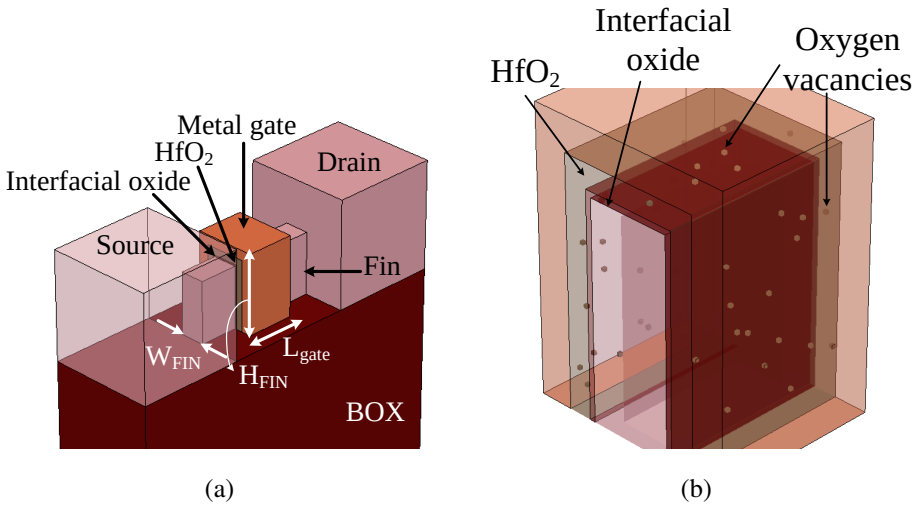



**Figure 11: (a) FinFET schematic. (b) Charged cubes (0.5 nm) representing a random placement and count of positively charged OV in the $HfO_2$ layer to induce local variability.**

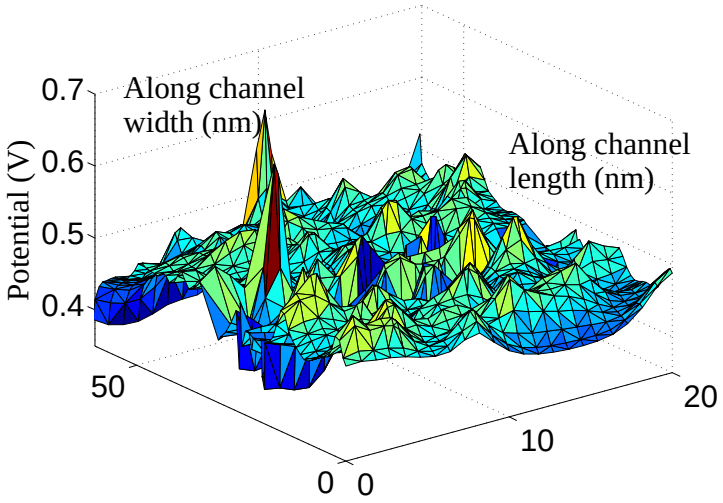



**Figure 12: OV-induced local variability in surface potential at the channel/oxide interface.**

### 3.1.3 Oxygen Vacancy-induced Variability in FinFET

The implications of oxygen-vacancy induced variability are considered on the non-traditional MOSFET structure, the FinFET. A schematic drawing of a FinFET was shown in Figure 11a. The gate-stack consists of interfacial oxide, $HfO_2$, and TiN metal electrode. The channel is considered to be undoped. FinFETs of 22, 16, and 11nm channel length technology are studied. Various geometry and electrical specifications of the structures are the same as in [57]. The traditional random-dopant-fluctuations (RDF), metal-gate-granularity (MGG), and fin/gate-edge-roughness (FER/GER) variability was studied in [57]. To facilitate comparison between the OV-induced variability and the other sources of variability, similar FinFET structures are being studied here.

An ensemble of $I_{DS} - V_{GS}$ characteristics ($V_{DS} = V_{DD} = 1$V) for the 22-nm technology n-FinFET with OVs is shown in Figure 13. Hereafter, the presented statistical characteristics are extracted over 100 random samples. By reducing the effective WF of the metal electrode, OVs shift the $V_{th}$ by ~0.17 V. The distribution of OV-induced $V_{th}$ variability is shown in Figure 13b, where OV induces ~20mV $3\sigma$-variability in $V_{th}$. In Figure 13c,

the correlation of $V_{th}$ quantiles against the standard Normal is shown for the various technologies. The $V_{th}$ distribution closely resembles the standard Normal. Moreover, the $V_{th}$ variability increases for the smaller channel length technology.

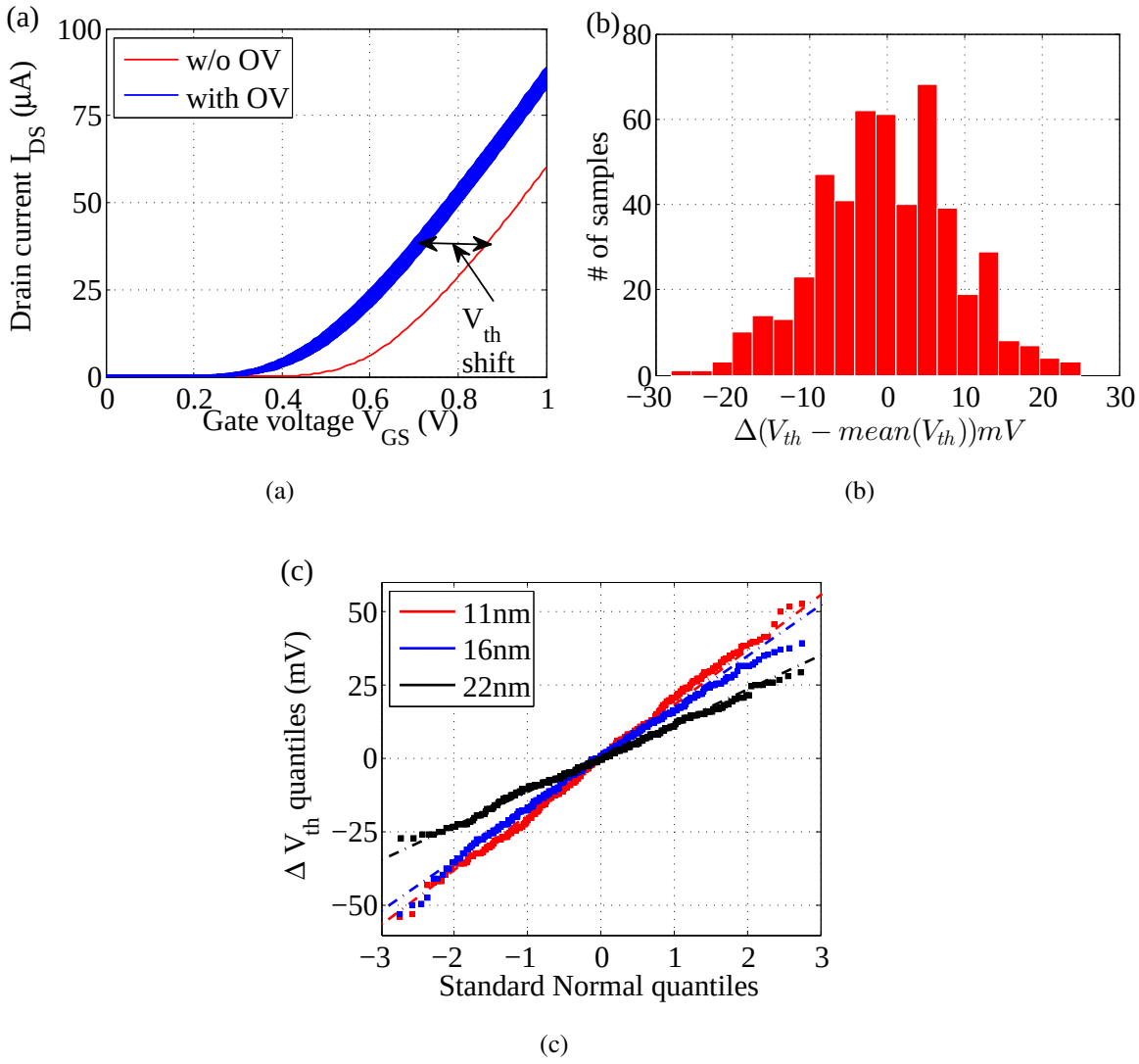



**Figure 13: (a) OV-induced $V_{th}$ shift and variability in $I_{DS}$-$V_{GS}$ characteristics of 22-nm FinFET. (b) Histogram of $V_{th}$ variability in 22-nm FinFET. (c) Correlation of OV-induced $V_{th}$ quantiles against the standard normal across technologies.**

The total variability of FinFET, consisting of RDF, MGG, FER, GER, and OV-induced variability, is shown in Figure 14. The other variability results (i.e., for RDF, MGG, FER,

and GER) were presented for similar FinFET structures in [57]. Here, the total variability increases for smaller channel length technologies, along with all the other variability components. By comparison with the other sources of variability, it is evident that the OV-induced variability becomes a significant component.

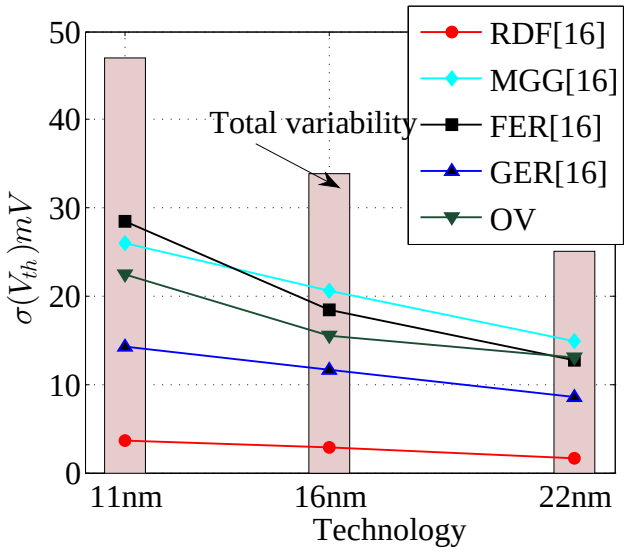
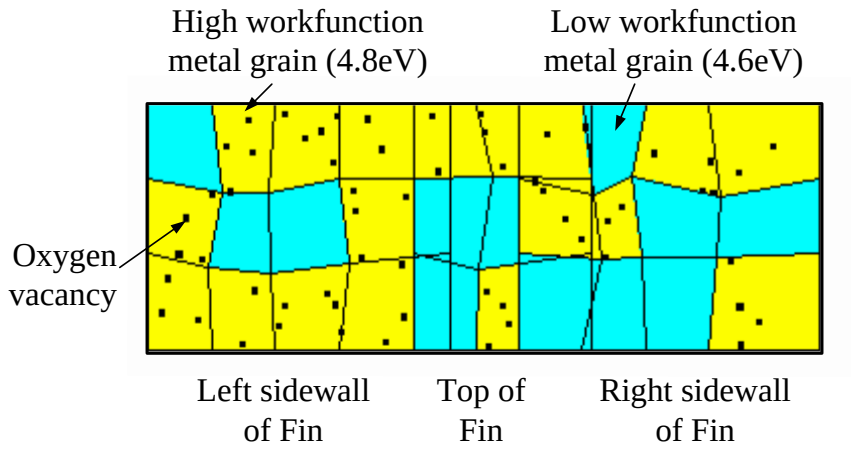


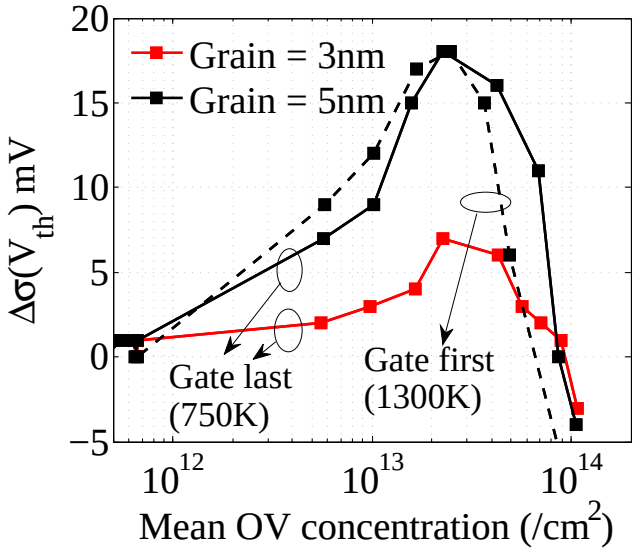

**Figure 14: OV-induced variability compared with the other sources of variability: RDF, MGG, FER, and GER.**

### 3.1.4 Implications of Metal Grain Granularity to Oxygen Vacancy Distribution

Notably, oxygen vacancies can also suppress the threshold voltage variability due to metal grain granularity (MGG). In a granulated metal gate, the oxygen vacancies are attracted towards a higher workfunction grain as shown in Figure 15a. The presence of a greater oxygen vacancy concentration underneath a high workfunction grain mitigates the surface potential variability due to MGG. In Figure 15b, the threshold variability reduction $\Delta V_{th}$ is shown at varying oxygen partial pressure conditions and for various average grain sizes. A variation in oxygen partial pressure affects the net concentration of oxygen vacancy as seen in Equation 10. Therefore, with an optimal oxygen vacancy placement, the effect of MGG induced variability is significantly suppressed.




(a)




(b)

**Figure 15: (a) Under metal gate granularity, a higher oxygen vacancy concentration under higher workfunction metal grain. (b) MGG-induced variability suppression at varying oxygen vacancy concentration.**

Therefore, while a high-$\kappa$ metal-gate process is useful to suppress gate-leakage, poly-depletion, Fermi-level pinning, and high gate-resistance, OV induced variability becomes a critical concern with this non-traditional gate-stack. Furthermore, OV induced variability becomes a greater concern than the other traditional (such as RDF) and emerging (such as MGG) variability mechanisms.

## 3.2 Threshold Voltage Modulation in Non-traditional Three-Dimensional MOSFET Integration

The three-dimensional (3-D) integration of MOSFETs minimizes interconnect power-dissipation and enhances bandwidth. This section investigates a physical phenomenon, i.e., the effect of the throughoxide-via (TOV) potential on the neighboring devices in a 3-D system. Considering a fully depleted silicon-on-insulator (FDSOI) transistor as the device-under-test, TOV induced critical challenges in a 3-D integrated FDSOI transistor are examined.

### 3.2.1 Through-Oxide-Via-induced Back-Gate Effect

TOVs in a 3-D system carry static (supply and ground) and dynamic (clock and logic signals) voltages [58]. In a 3-D system, TOVs are isolated from SOI devices through the surrounding dielectric. The static or dynamic voltage in TOVs modulates electric field in the surrounding dielectric [see Figure 16]. The net effect is a variation in the back-gate potential of the neighboring SOI devices. For FDSOI devices, a variation in the backgate bias modulates device electrostatics and, hence, electrical characteristics.

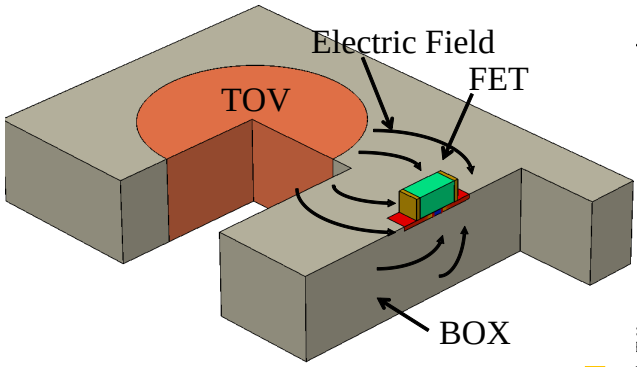



**Figure 16: Interaction of TOV and FD-SOI device in a 3-D integrated system.**

### 3.2.2 Through-Oxide-Via-induced On/Off-current and Threshold Voltage Modulation

The threshold voltage of a FDSOI device strongly depends on the back-gate bias [59]. Hence, the TOV potential variation strongly modulates the threshold voltage of the devices.

For a closely located FET, a higher TOV potential results in an increased (positive) back-gate potential for the FDSOI device. A higher back-gate potential reduces the threshold voltage of the NFET. On the other hand, a lower TOV potential decreases the back-gate potential, resulting in lower effective gate-to-back-gate potential $V_{GB}$. A lower $V_{GB}$ value thus reduces the threshold voltage of the PFET.

Since the variation in the threshold voltage of the FET is due to the TOV-induced back-gate effect, it follows similar geometrical dependence as described in the earlier section. Figures 17 & 18 show the change in the threshold voltage for the NFET and the PFET when the via potential changes from 0 to 1 V (VDD). The change in the threshold voltage is shown for different separations between the TOV and the FET and for varying diameters of the TOV. A reduction in the threshold voltage exponentially increases the subthreshold leakage. Therefore, in 3-D systems, when the TOV potential is high, it increases the leakage of the NFET, whereas, when the TOV potential is low, it increases the leakage of the PFET. Figures 17 shows the increase in the leakage of the NFET for the TOV potential at VDD from the leakage for the TOV potential of 0 V (top). In Figure 18, PFET leakage increases for the TOV potential at 0V from the leakage for the TOV potential of VDD. A significant variation in the leakage current is evident. For example, for a 5-$\mu$m TOV, both NFETs and PFETs as far as 8 $\mu$m can experience almost ~50%100% increase in the leakage, depending on the TOV potential. Corresponding variations to the on-current are also analyzed. However, since the oncurrent has a weaker dependence on the threshold voltage, the corresponding variations is smaller (~0.5%2.5%).

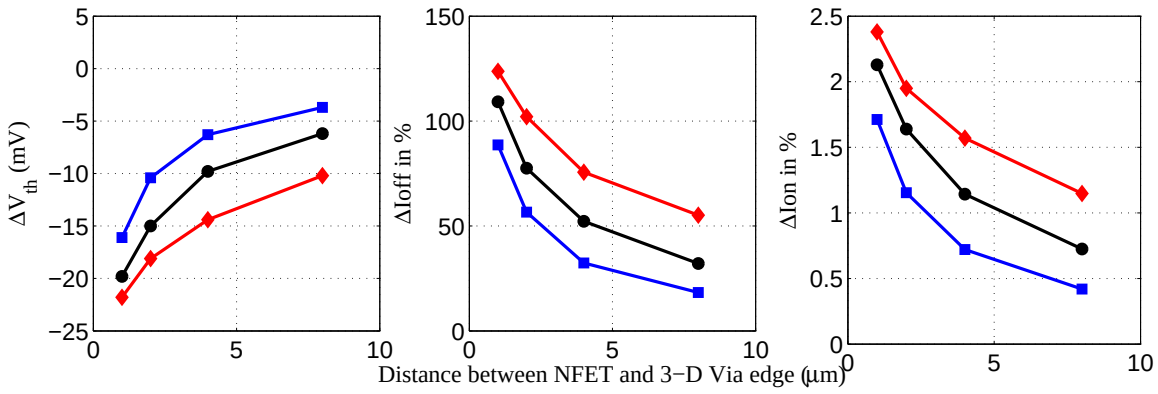



**Figure 17: Effect of the TOV on the NFET: (a) threshold voltage, (b) subthreshold leakage $I_{off}$ , and (c) on-current ($I_{on}$). [TOV diameter, blue: 1 $\mu$m, black: 2 $\mu$m, red: 5 $\mu$m ]**

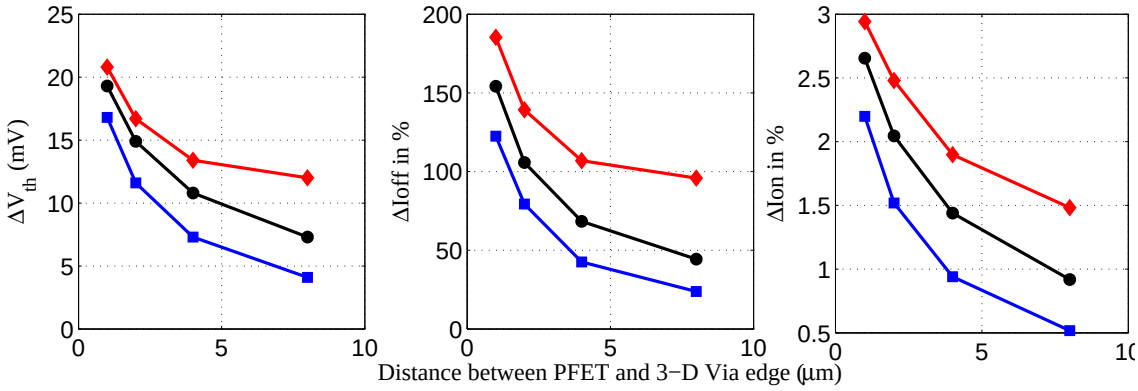



**Figure 18: Effect of the TOV on the PFET: (a) threshold voltage, (b) subthreshold leakage $I_{off}$ , and (c) on-current ($I_{on}$). [TOV diameter, blue: 1 $\mu$m, black: 2 $\mu$m, red: 5 $\mu$m ]**

## 3.3 Conclusions

This chapter has discussed the emerging variability sources in non-traditional MOSFET design and integration. Although, a high-$\kappa$/metal gate process in MOSFET suppresses gate-leakage, poly-depletion, and Fermi-pinning, a workfunction variability is introduced as a result of thermodynamic defect, oxygen-vacancy (OV). A 3-D integration of MOSFET is promising for suppressing interconnect power-dissipation and enhancing bandwidth, however, 3-D via induces on/off-current and threshold-voltage modulation. Therefore, along

with a limited switching slope, foregoing and emerging process induced variability limits VDD scaling and energy-efficiency of MOSFET-based designs. Therefore, it becomes critical to explore the computing devices which can overcome the switching slope limitations of MOSFET, and the computing architectures which are more resilient against process variation. The subsequent chapters in this dissertation explore TFET and TFET-based non-traditional computing architectures for the above goals.

# CHAPTER 4

# SINGLE NEURON-BASED POWER-GATING EFFICIENCY LEARNING, AND APPLICATION TO SELF-ADAPTIVE POWER-GATING

A low-power learner-circuit based on a single neuron-dynamics is presented in this chapter to characterize the trade-off between leakage-saving and transition-energy overhead in power-gating (PG). A self-adaptive PG scheme is demonstrated that utilizes the learner circuit to adaptively invoke PG only when leakage saving is more than the transition energy overhead. A 130nm test-chip demonstrates functionality of the learner-circuit and its application to adaptive-PG under varying process, temperature, and idle signal pattern. Various contributions discussed in this chapter were published in [60, 61].

## 4.1 Fine-Grained Power-Gating

Fine-grained PG is gaining interest to save leakage-energy during brief idle periods in individual power-domains. Various micro-architectural control signals such as block access signal for caches [62], clock-gating signals for cores [63], or input/output data-phases of look-up tables (LUTs) for FPGA [64] can be used to locally extract idle signal (IDL) for run-time fine-grained PG. However, the leakage saving ($\Delta P_{leak}$) through PG comes at the expense of transition-energy overhead ($E_{ov}$). When the overhead-energy exceeds the leakage-energy saving (i.e. $\int \Delta P_{leak} dt$), PG becomes energy-inefficient. The idle-signal (IDL) generated purely based on the logical activity does not account for the $\Delta P_{leak} E_{ov}$ trade-off, and hence, degrades the PG energy-efficiency.

Various PG schemes utilize 'PG controller' to extract the PG signal from IDL while accounting for $\Delta P_{leak} E_{ov}$ trade-off [65, 66, 67]. However, due to the process/temperature (P/T) dependence of leakage-current, $\Delta P_{leak}$ varies from chip-to-chip, within a chip, and over time. Hence, on-line sensing of local P/T variation is critical to accurately account for $\Delta P_{leak}$ $E_{ov}$ trade-off during PG signal generation. A small area/power overhead of the

PG-controllers is also desired.

## 4.2 Single Neuron Dynamics-based Power-Gating-Efficiency Learner

A low-power power-gating-efficiency (PGE) learner based on a single neuron-dynamics is discussed here to generate PG signal from IDL considering variations in process, temperature, and activity pattern of IDL. Consider a single neuron dynamics as

$$\frac{dx}{dt} = -\frac{x}{R} + \sum_i g(y_i) \tag{11a}$$

$$y = f_{th}(x). \tag{11b}$$

Here, '$x$' represents the state of the neuron, and '$y$' represents its output. The term '$-x/R$' represents the rate of loosing (forgetting) the current state of the neuron. And, the term '$g(y_i)$' represents the rate of learning from the other neurons (or, the external environment) to update its current state. '$f_{th}$' is a threshold-function to generate '$y$' from '$x$'.

When the neuron-dynamics is applied to learn PGE, a single neuron learns $E_{ov}$ in the current state (i.e., '$x$'). The neurons looses (forgets) its current-state by $\Delta P_{leak}$ rate to represent the net energy-saving, $\int \Delta P_{leak} dt - E_{ov}$, as its current-state. A comparator reading the current-state of PGE learner creates a binary output '$y$' to denote any positive leakage-energy savings. If $y = 1$ (i.e., positive leakage-energy savings), the PG signal for a domain follows IDL, otherwise, the PG is not employed on the domain.

## 4.3 Power-Gating-Efficiency Learner Implementation and Operation

### 4.3.1 Power-Gating-Efficiency Learner Implementation

Consider the PGE learner design for a power-gated system using a p-type PG transistor (as shown in Figure 19). The schematic of the PGE learner is shown in Figure 20. The learner is controlled by the idle-signal of the PG-domain (i.e., IDL). A 0-to-1 transition in IDL shifts a PG domain to the idle-mode. The mode-transition incurs energy-overhead ($E_{ov/tran}$) as a result of switching of the virtual VDD, intermediate logic-nodes, and gate-capacitance of the sleep-PMOS. Power gated mode (IDL=1) reduces leakage-power by $\Delta P_{leak}$. With

a 0-to-1 transition in IDL, transistor $TX_N$ is activated through the edge-detector and the overhead-expense is learned by a drop in the potential of the node $ST$.

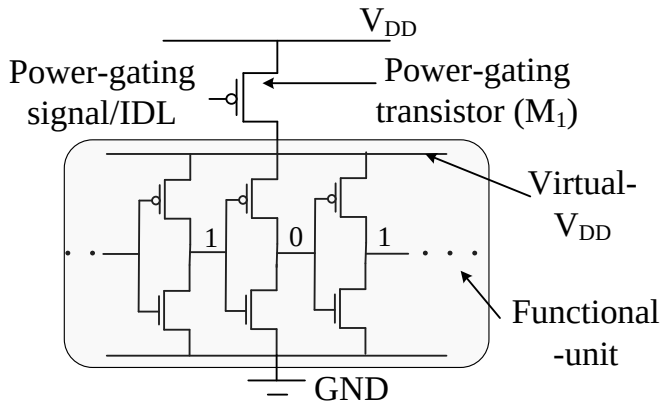



**Figure 19: A demonstration of power-gating.**

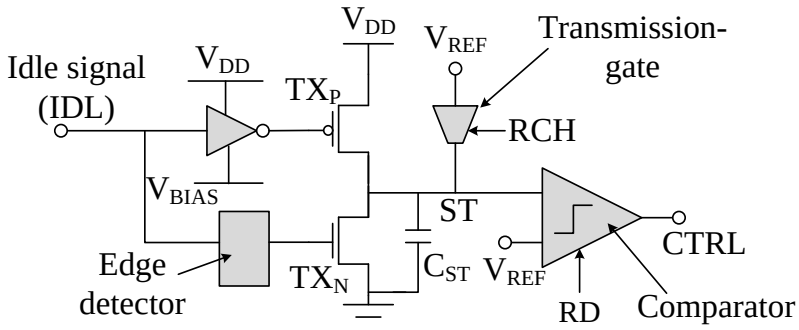



**Figure 20: Power-gating-efficiency learner schematic.**

When IDL=1, the sub-threshold biased transistor $TX_P$ increases the potential of node $ST$, and the record of previous overhead-expense is gradually erased (forgotten). $TX_N$ and $TX_P$ can be designed to ensure that $V(ST)$ follows $E_{ov/tran}$ and $\Delta P_{leak}$, respectively, of the PG-domain. This is achieved through the design equations,

$$I(TX_P) \propto \Delta P_{leak} \tag{12a}$$

$$T_{PW} \times I(TX_N) \propto E_{ov/tran} \tag{12b}$$

$$\frac{I(TX_P)}{T_{PW} \times I(TX_N)} = \frac{\Delta P_{leak}}{E_{ov/tran}} = \frac{VDD \times (I_{NPG} - I_{PG})}{E_{ov/tran}}, \tag{12c}$$

where $I(TX_P)$ is the current through $TX_P$, $T_{PW}$ is the pulse-width generated by the edge-detector, and $I(TX_N)$ is the on-current of $TX_N$. $I_{NPG}$ and $I_{PG}$ are the leakage-currents of the PG-domain in the non-power-gated and power-gated modes, respectively. To minimize the area of PGE-learner, $TX_N$ can be of the minimum-width, and $TX_P$ can be determined by Equation 12c. The design equation (Equation 12c) is required to also track the temperature changes. $E_{ov/tran}$ is relatively insensitive to temperature, while $\Delta P_{leak}$ is very sensitive to temperature. Thus, temperature sensitivity of $I(TX_P)$ can be matched to track $\Delta P_{leak}$ under temperature variations. In the subthreshold-region, the temperature sensitivity of $I(TX_P)$ is given by

$$\frac{1}{I(TX_P)}\frac{dI(TX_P)}{dT} \propto V_{th,TXP} - V_{SG}. \tag{13}$$

The sensitivity is higher at a smaller source-gate voltage ($V_{SG}$), and appropriate $V_{BIAS}$ and $TX_P$ width can be designed. The operation of PGE learner is independent of the capacitance ($C_{ST}$) at the node $ST$ [see Equation 12c]. However, a very small $C_{ST}$ can result in saturation of the node $ST$ within the observation period (i.e. $V(ST) \rightarrow$ VDD or zero). On the other hand, a larger $C_{ST}$ reduces the sensitivity of $TX_P$ & $TX_N$ induced modulation to $V(ST)$. Hence, a larger $C_{ST}$ can increase the risk of erroneous comparator output in the presence of comparator-off-set and/or dynamic noise. Hence, an appropriate $C_{ST}$ needs to be used considering these trades-off.

### 4.3.2 Break-Even and Power-Gating-Efficiency Tracking

The operation waveform of PGE-learner are shown in Figure 21. The PGE-learner is used to predict the process/temperature dependent break-even time. The break-even (BE) time is defined as the minimum-off period ($T_{OFF}$) that makes PG beneficial i.e. $\int \Delta P_{leak} dt > E_{ov}$. In Figure 21a, following 0→1 transition in IDL signal, $TX_N$ reduces $V(ST)$ which is slowly raised again by $TX_P$ at IDL=1. At the time when $V(ST)$ returns back to its initial condition (i.e. $V_{REF}$), the comparator's output (CTRL) changes from 0→1. Following Equation 12c, at this time $\int \Delta P_{leak} dt > E_{ov}$. Hence, BE of the PG-domain is estimated by this duration as

shown in Figure 21a.

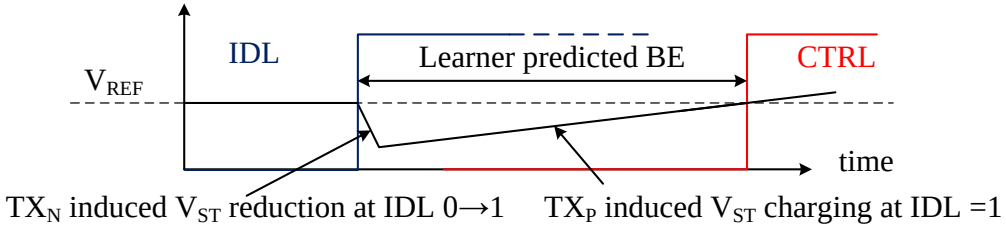
The PGE-learner tracks the net energy-savings over an extended observation period as shown in Figure 21b. During the total observation period ($T_{obs}$), the energy-saving in the PG-domain can be expressed as

$$E_{leakage,saving} = \int \Delta P_{leak} dt - E_{ov} = T_{OFF} \Delta P_{leak} - N \times E_{ov/tran}, \tag{14}$$

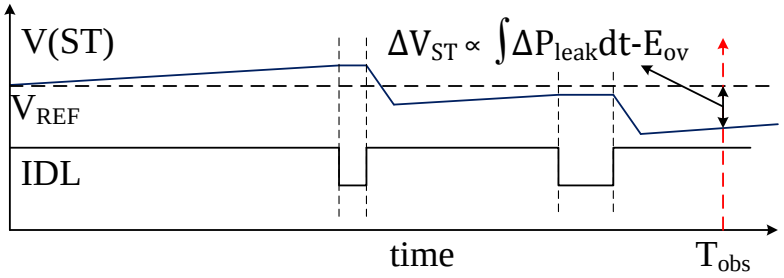
where $T_{OFF}$ is the total off period (i.e. with IDL=1) and $N$ is the total numbers of 0→1 transition in IDL during $T_{obs}$. At the end of $T_{obs}$, following (Equation 12c) the change in potential $V(ST)$ is:

$$\Delta V_{ST} = V_{ST}(T_{obs}) - V_{REF} \propto E_{leakage,saving}. \tag{15}$$

Hence, $\Delta V_{ST} > 0$ indicates the net energy-saving and beneficial PG during $T_{obs}$. On the other hand, $\Delta V_{ST} < 0$ denotes excessive overheads. RD enabled comparator at the end of $T_{obs}$ reads the polarity of $\Delta V_{ST}$ as the digitized output (CTRL).




(a)




(b)

**Figure 21: Power-gating-efficiency learner operation: (a) break-even time tracking and (b) leakage-energy savings $\int \Delta P_{leak} dt - E_{ov}$ tracking.**

## 4.4 Self-Adaptive Power-Gating with Power-Gating-Efficiency Learner

A self-adaptive PG (SAPG) scheme as shown in Figure 22 utilizes the PGE learner's output to minimize the leakage energy. In Figure 20, RD determines the observation cycle ($C_N$) of the PGE-learner to analyze the activity pattern (i.e. IDL). At the beginning of $C_N$, the node ST is charged to the reference voltage ($V_{REF}$) through the transmission gate (in Figure 20) using the control signal RCH. At the end of $C_N$, $V_{ST}$ is compared to $V_{REF}$ using a clocked-comparator (in Figure 20) controlled by the signal RD, and generates the output signal CTRL. If CTRL=0 ($\int \Delta P_{leak} dt < E_{ov}$ case), the multiplexer in Figure 22 avoids PG in the next cycle $C_{N+1}$. Hence, PG is only invoked if it was beneficial in the previous observation cycle. Since the observation cycle is used to learn the PG efficiency, it is referred to as the learning cycle.

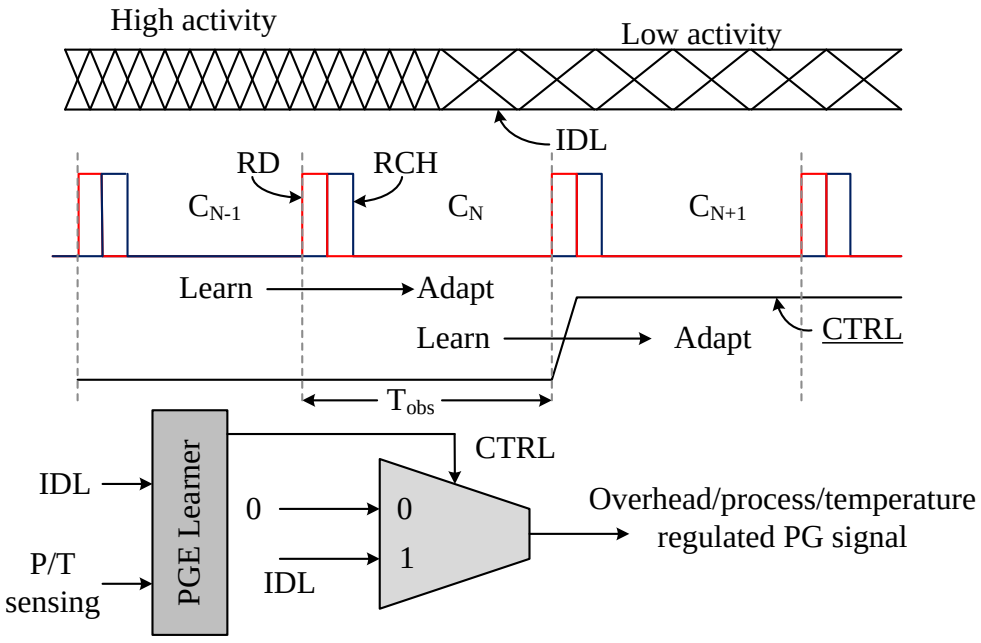



**Figure 22: A demonstration of self-adaptive power-gating.**

## 4.5 Test-chip and Measurement Results

The PGE learner is demonstrated in a test-chip designed in 130nm CMOS (Figure 23, Table 1). An internal signal-generator generates the RD/RCH signals for learner. The

IDL signals were generated internally or supplied externally through field-programmable-gate-array (FPGA). Four PG domains were implemented with regular-$V_{th}$ (RVT) and low-$V_{th}$ (LVT) transistors. Each domain consisted of a chain of 301 inverters to emulate a very fine-grain PG condition. PGE-learners were embedded within each domain. The RVT and LVT PG-domains emulate extreme within-chip process variations. To emulate dynamic temperature variations a heater, designed with diffusion resistors, is embedded in the design. Heater-power is varied to control on-chip temperature.

**Table 1: Proto-type chip specifications**

| | |
|---|---|
| Technology | 130nm |
| VDD | 1.2V |
| PG-transistor area | $120\mu m^2$ ( 7% of the block area) |
| PGE-learner area | $180\mu m^2$ |
| Block-leakage | 84nA(RVT), 560nA(LVT) at room-temperature |

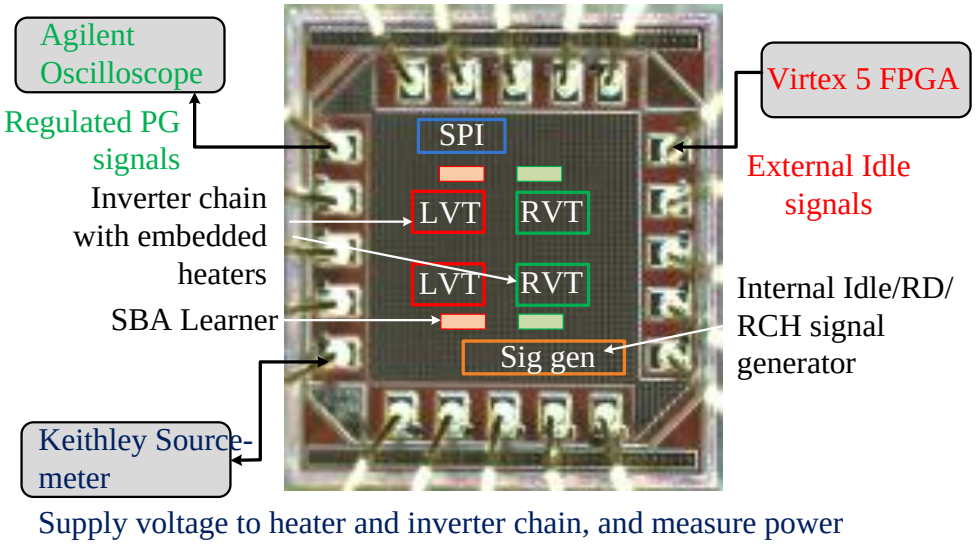
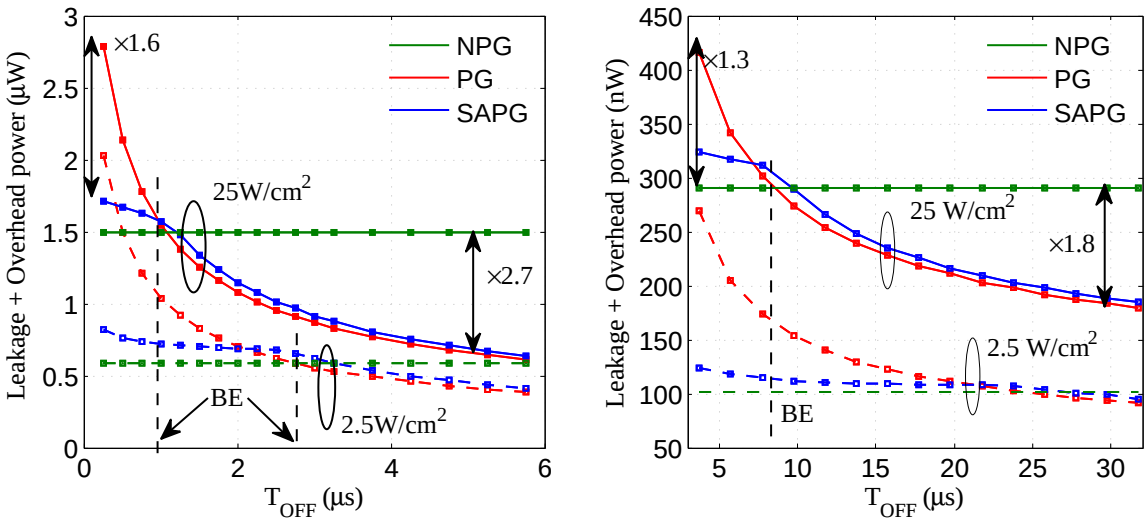



**Figure 23: Test-chip photo-shot and experimental setup.**

### 4.5.1 Characterization of Break-Even Learning Accuracy

The accuracy of the proposed PGE learner is characterized by comparing the actual BE of a PG domain with the BE predicted by the learner [Figure 24]. The actual BE is measured by directly power-gating the domain with the periodic idle patterns (IDL) of fixed on time ($T_{ON}$ = 250ns) and varying off time ($T_{OFF}$). At a lower $T_{OFF}$, because of the lesser net off-time and higher overheads, the average total (leakage + overhead) energy of the domain increases. The corresponding results (label: PG) are shown in Figure 24. The non power gated (label: NPG) case is the leakage power in absence of PG. The measured (or actual) BE of PG domain is defined as the $T_{OFF}$ at which the NPG and PG curves intersects (Figure 24). For an idle pattern with $T_{OFF}$ = BE, the overhead is equal to the leakage saving. The measured BEs for the LVT design is $1\mu$s and $2.7\mu$s for 25W/cm$^2$ and 2.5W/cm$^2$ heater powers, respectively. As a result of to its lower leakage, the RVT design exhibits a larger BE time ( $27\mu$s) as shown in Figure 24b.




(a) (b)

**Figure 24: Measurement results with idle signals of varying $T_{OFF}$ and comparison of power in various modes: (a) LVT design and (b) RVT design.**

The BE predicted by the PGE learner is measured by applying the similar IDL signal of varying $T_{OFF}$ and observing the output PG signal. The measured waveforms in Figure 25 show that for the learner blocks the IDL patterns with very low $T_{OFF}$ (i.e. the $\int \Delta P_{leak} dt < E_{ov}$ case resulting in CTRL=0). The IDL patterns with higher $T_{OFF}$ pass through (i.e. PG = IDL) the learner. The learner's predicted BE is defined as the $T_{OFF}$ of IDL when CTRL makes the 0→1 transition (i.e. the PG signal starts to follow IDL). The heater power (or temperature) dependent BE tracking is also shown in the figure. At the lower heater power (or lower temperature) the learner correctly identifies the BE at a higher $T_{OFF}$ value.

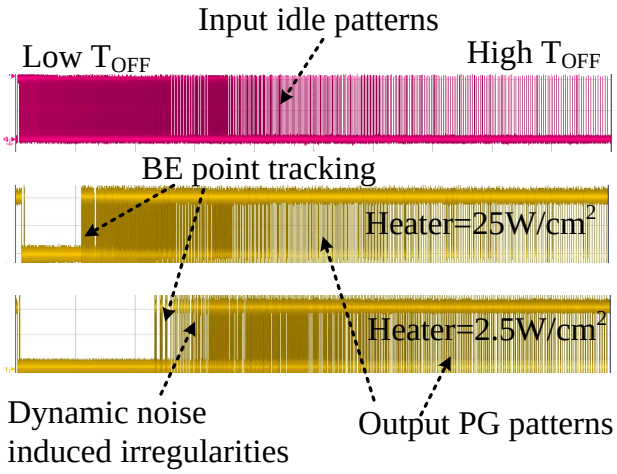



**Figure 25: A Measurement results on break-even point tracking at varying on-chip temperature.**

### 4.5.2 Self-Adaptive Power-Gating at varying Process/Temperature Conditions

The effectiveness of SAPG is evaluated for varying activity of the IDL signal, process, and temperature conditions. The power dissipation (leakage + overhead + learner) of SAPG (label: SAPG) is compared with that of the PG and NPG cases. First, the effectiveness of SAPG is studied at varying activity (i.e. varying $T_{OFF}$) of IDL signal (Figure 24). At very low $T_{OFF}$, the SAPG reduces power by avoiding PG, while at very high $T_{OFF}$ SAPG reduces power by performing PG. The effectiveness of SAPG is next studied for constant activity ($T_{OFF} = 2\mu s$) of the IDL signal but varying heater power and temperature. As

shown in Figure 26, PG becomes expensive for heater power < 10W/cm$^2$ and the SAPG configures the LVT domain to the NPG mode. Therefore, the SAPG successfully optimally configures a PG domain to the lower power mode between PG and NPG.

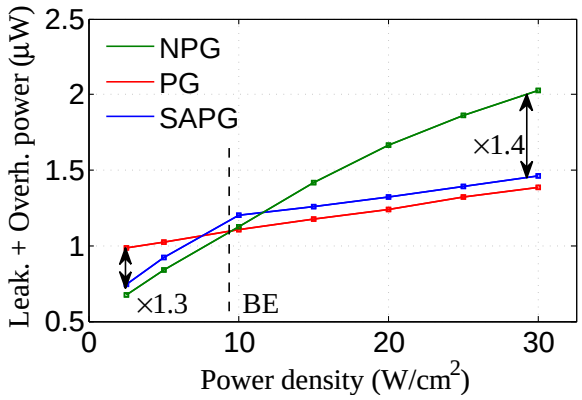



**Figure 26: Self-adaptive power-gating across power-density ($T_{OFF} = 2\mu s$), LVT design.**

### 4.5.3 Implications of the Learning Cycle

Implications of the learning cycle in SAPG are discussed in Figure 27. The proposed SAPG dynamically adapts to the varying PG patterns as illustrated in the example waveform in Figure 27a. However, the effectiveness of the adaptation depends on the learning cycle (i.e. the period of RD). If there is a drastic change in the activity pattern, a larger learning period ($T_{learn}$) suffers from the lag in adaptation and reduces the effectiveness of SAPG. The effect of adaptation lag is illustrated in Figure 27b which shows the unwanted PG at high IDL activity and missed PG opportunities at the low IDL activity. A smaller learning cycle reduces the missed opportunity and unwanted PG regions. However, the smaller learning cycle has larger-power overheads because of frequent access to the comparator and the recharge circuit. A pseudo-random-sequence (PRS) of varying $T_{OFF}$ is considered to study the effect of $T_{learn}$. The power with the SAPG is measured considering the PRS idle pattern and a periodic idle pattern at varying $T_{learn}$ (Figure 27c, the power of the each case is normalized with respect to their values at $T_{learn} = 250\mu s$). While for a periodic pattern larger $T_{learn}$ is advantageous, for the PRS case reducing $T_{learn}$ initially improves SAPG (the

optimal $T_{learn}$ for this example is $\sim 7\mu s$). As the functionality of SAPG does not depend on $T_{learn}$, it can be programmed on-line depending on the application.

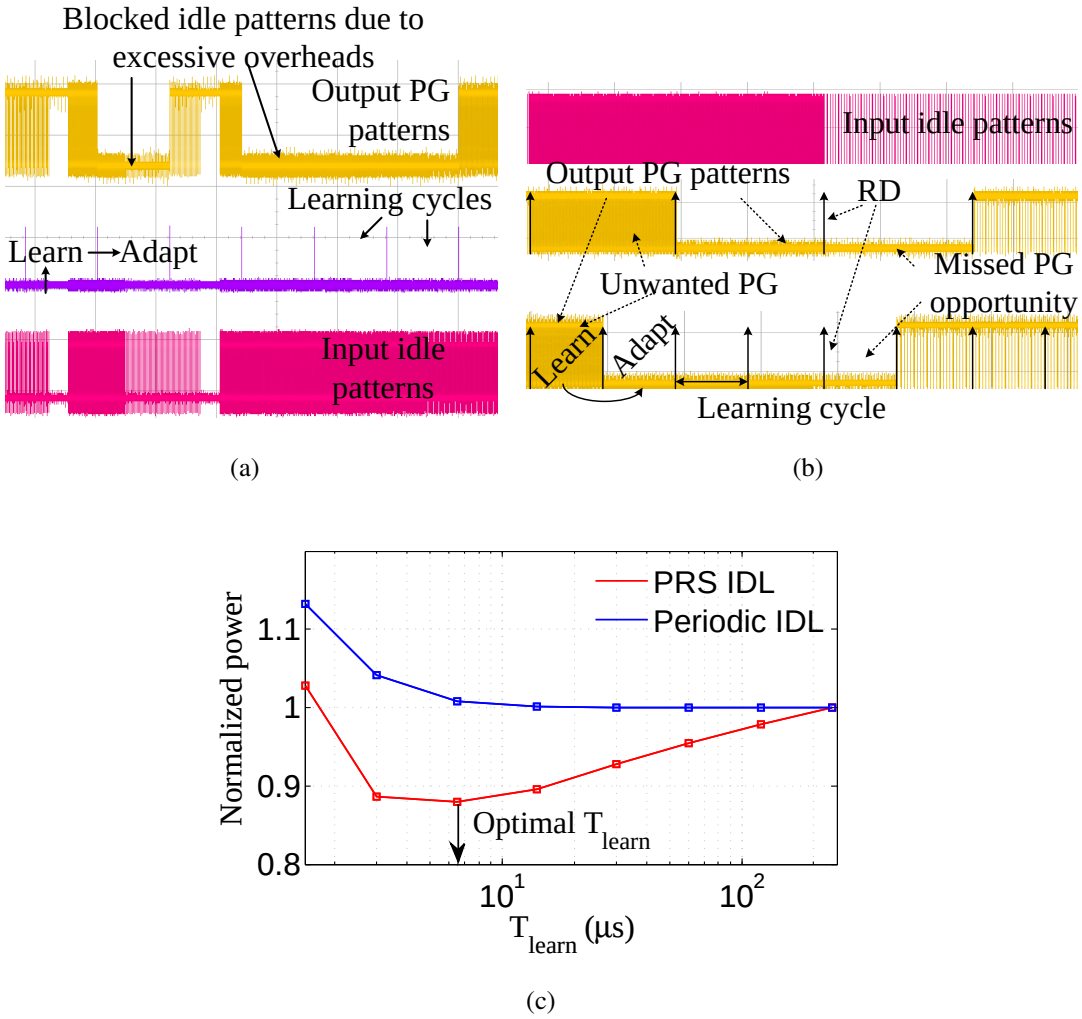



**Figure 27: Self-adaptive power-gating (SAPG) for varying activity patterns: (a) measured waveform showing SAPG at randomly varying activity, (b) measured waveform showing the better adaptation at smaller learning cycle, and (c) measured total (leakage + overhead + leaner) power at various learning cycle for different patterns.**

## 4.6 Limitations of the PGE Learner-based Power-Gating

The key challenge for the PGE learner is the prediction inaccuracies caused by the comparator's offset and dynamic-noise. Especially, at the condition $\int \Delta P_{leak} dt \sim E_{ov}$, the developed $V(ST)$ in a PGE-learner is close to $V_{REF}$ and hence, dynamic-noise and offset in the comparator affects the transition of the CTRL signal as shown in Figure 28.

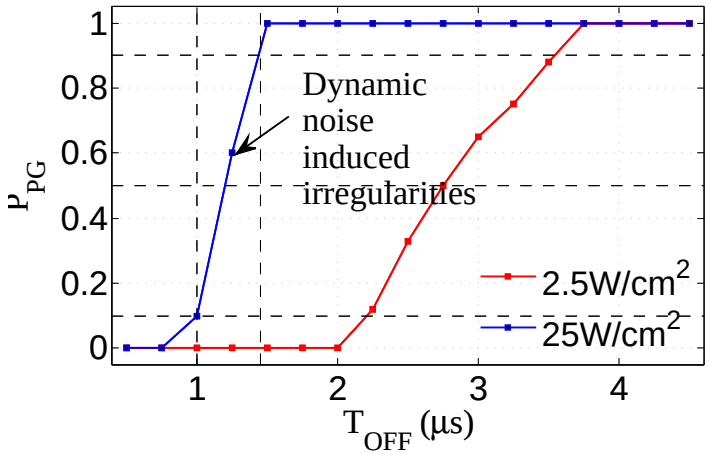
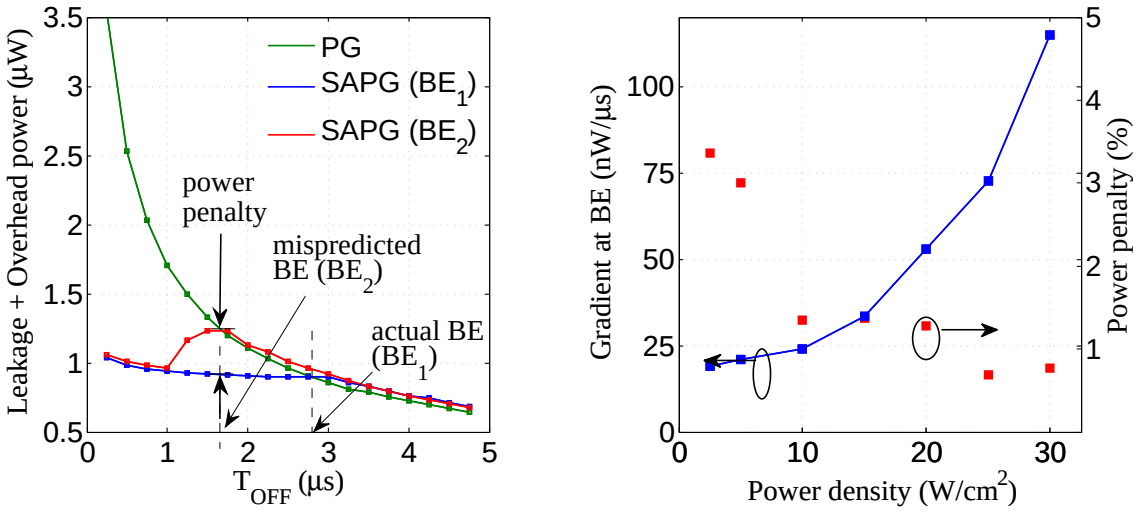



**Figure 28: PGE Learner's probability to power-gate under dynamic noise.**

The BE misprediction due to the dynamic-noise/offset leads to inaccurate transitions between the PG and NPG modes [Figure 29]. The effect of misprediction is illustrated in Figure 29a which shows that the SAPG power with mispredicted BE (=$BE_2$) can be higher than the power with the ideal prediction ($BE_1$). The maximum penalty due to the misprediction is measured for various heater-power (i.e., on-chip temperature) but at a constant activity ($T_{OFF}$) of IDL. At a lower on-chip temperature (lower heater-power), the effect of dynamic-noise is higher because of the lower current through $TX_P$ and a higher impedance at the node $ST$ [Figure 28]. However, it can be interpreted from Figure 24 that at low-temperature, the leakage + overhead power of a PG-domain changes only slowly with $T_{OFF}$ near the BE point. The measurement of the gradient-of-power against $T_{OFF}$ at BE for different heater-power verifies the above observation (Figure 29b). Hence, even a large inaccuracy in the BE causes less power-penalty at a lower temperature. Hence, even

assuming a worst case of misprediction, the power-penalty over a wide range of on-chip power-density remains within 5% in Figure 29b.




(a) (b)

**Figure 29: Implications of PGE learner-inaccuracy: (a) power-penalty in SAPG due to BE inaccuracy, and (b) power-penalty due to dynamic-noise across power densities (temperatures).**

Furthermore, dedicated PGE learners for several PG-domains increase implementation-complexity. The analog-voltages ($V_{REF}$, $V_{BIAS}$) and digital-signals (RD, RCH) are required for each PGE-learner's operation. Therefore, sharing $V_{REF}$, $V_{BIAS}$, RD, and RCH generation circuits among multiple PG-domains is necessary to reduce this overhead. Nonetheless, fine-grained distribution of these voltages and signals will add to the routing overheads, and a full-chip analysis will be required to optimize the number/placements of the learners. Also, note that the discussed inverter-chain-based logic-block ignores the input vectors and internal-state dependent leakage-variation in a logic-block. The state-dependent leakage of the PG block can be considered in the PGE learner by programming the width of $TX_P$ based on input/output flip-flop states before each PG event. However, as leakage of a logic-block experiences only limited change with the input vector [68], such state-dependent inaccuracy of the learner can be ignored to avoid a more complex PGE-learner implementation.

## 4.7 Comparison with the Prior Works

A comparison of the proposed PGE-learner with the existing PG-controllers [] is presented in Table 2. Since the detailed controller schematics were not presented in the works [65, 66, 67], a standard implementation with D-flip-flops, counters, and state-machine is assumed while estimating the transistor count. For [67], a look-up-table (LUT) of eight-words with eight-bits is assumed. The primary advantage of the proposed neuron dynamics-based design is a small-area/power as a result of the elimination of counters, registers, state-machines, and LUTs. The proposed learner also only dissipates ~100nW power (in 130nm CMOS) which is much lower than the other works: 48.3W (in 90nm technology [65]) and 184W (in 90nm technology [66]). The asynchronous implementation of the controller in [64] has an equivalent area, but it lacks adaptation against process/temperature variation and history.

**Table 2: Comparison of PGE-learner with the prior-art**

| Properties | PGE learner | [65] | [66] | [64] | [67, 69] |
|---|---|---|---|---|---|
| No. of transistors | 57 | 86 | 210 | 500 | 600 |
| Area | Small | Small | Medium | Large | Large |
| Power | Small | Small | Large | Large | Large |
| Proc./Temp. based adaptation | Yes | No | No | No | Yes |
| History based adaptation | Yes | No | No | Yes | Yes |

## 4.8 Conclusions

A PGE learner-based on the dynamic of a single neuron operation was discussed in this chapter. The PGE learner enables dynamic configuration of a circuit domain to PG and NPG mode to optimally trade-off leakage-saving with energy-overhead. Measurement-results in 130nm CMOS demonstrate that the learner accurately tracks the break-even cycle

of a PG-domain and performs self-adaptive power-gating. The leaner has lower area/power overhead than the existing techniques while enabling process, temperature, and history based adaptation. Therefore, the proposed leaner can be distributed locally to facilitate fine-grain power-gating and reduce leakage of digital-system.

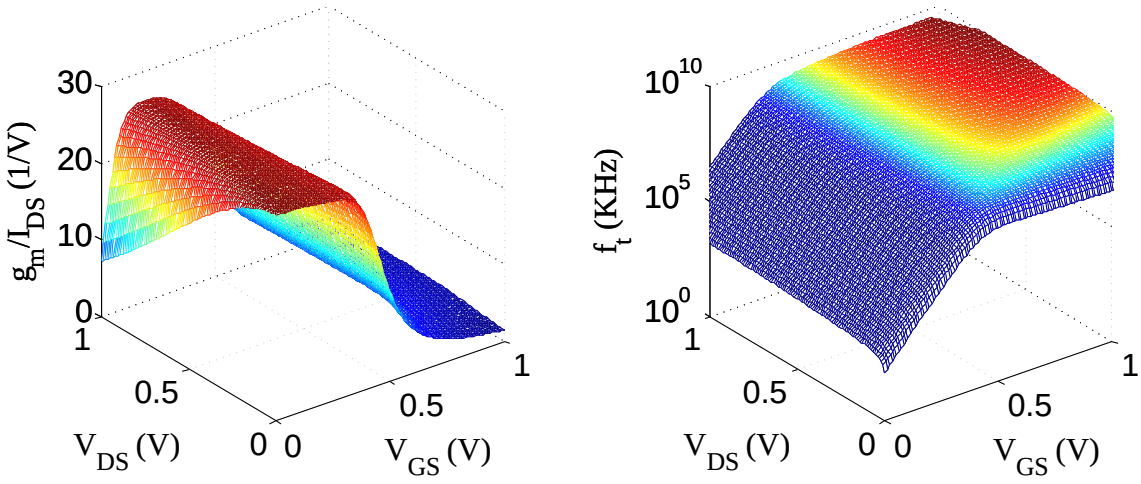
# CHAPTER 5

# TUNNELING-FIELD-EFFECT-TRANSISTORS TO ENABLE ULTRA-LOW-POWER ANALOG-COMPUTING

The potential of Tunneling-Field-Effect-Transistors (Tunnel FETs or TFETs) is discussed in this chapter to enable ultra-low-power analog-designs. TFET can achieve ultra-low quiescent current (<pA) and higher transconductance-per-bias-current than the conventional transistors MOSFET, and a very high output-resistance to enable low-power, energy-efficient analog-circuits. Specifically, challenges and opportunities in the design of TFET-based Operational-Transconductance-Amplifier (OTA) is discussed in this chapter. An ultra-low-power, energy-efficient TFET-OTA becomes the integral component in TFET-based large-scale, energy-efficient analog neuromorphic-computing as discussed in the successive chapters. Various contributions discussed in this chapter were published in [70, 71].

## 5.1 Motivation to Employ TFET for Low-Power Analog-Computing

To illustrate the potential of TFET for low-power analog-computing, let's first consider the analog-operation of the conventional transistor MOSFET. Transconductance-per-bias-current ($g_m/I_{DS}$) and transition-frequency ($f_T$) of N-MOSFET, in 90nm CMOS technology, is demonstrated across bias conditions in Figure 30. It is observed that $g_m/I_{DS}$ is significantly enhanced in the subthreshold-region (i.e., $V_{GS} < V_{th}$, $V_{th}$ is the threshold-voltage of the transistor). An enhanced $g_m/I_{DS}$ in the subthreshold-region is a result of the exponential dependence of drain-current to gate-voltage in the transistor. However, because of a lower drain-current, $f_T$ of the transistor reduces in the subthreshold-region. Hence, when the speed requirements are not stringent, the subthreshold-operation of MOSFET will be the most energy-efficient for analog-designs.




(a) (b)

**Figure 30: Bias-dependences for n-MOSFET in 90nm CMOS: (a) $g_m/I_{DS}$ and (b) $f_T$. [$V_{th}$ = 0.4V]**

In the subthreshold-region, the switching-slope (SS) of transistor relates to its transconductance ($g_m$) as

$$\frac{1}{SS} = \frac{\Delta log_{10} I_{DS}}{\Delta V_{GS}} = \frac{1}{log(10)} \frac{dlog(I_{DS})}{dV_{GS}} = \frac{1}{log(10)} \frac{1}{I_{DS}} \frac{dI_{DS}}{dV_{GS}}, \tag{16a}$$

$$\frac{g_m}{I_{DS}} = \frac{log(10)}{SS}. \tag{16b}$$

Hence, an even higher $g_m/I_{DS}$ in MOSFET is constrained by its SS. Meanwhile, SS in MOSFET is bounded by the thermal-energy of its charge-carriers ($kT$), and SS < 60mV/decade (at room temperature). TFET, however, uses quantum-mechanical band-to-band-tunneling (BTBT) for its charge-conduction, and unlike MOSFET, SS in TFET is not bounded by the thermal-energy of carriers. TFETs have shown to achieve SS lower than 60mV/decade [72]. With a lower SS, therefore, TFET can achieve higher $g_m/I_{DS}$, and TFET-based analog-designs can be more energy-efficient than MOSFET.

## 5.2 TFET Characteristics

Band-to-band-tunneling (BTBT) based charge-conduction in TFET was discussed in Chapter 2. The role of channel-material engineering and a source/channel heterojunction was also discussed to overcome the on-current limitations in silicon-TFET. In this section, simulation-methodology to evaluate TFET is presented, and key-characteristics of TFETs of various channel-materials are compared. For CMOS compatibility, silicon/germanium heterojunction TFET is considered. For higher performance, GaSb/InAs heterojunction TFET is considered.

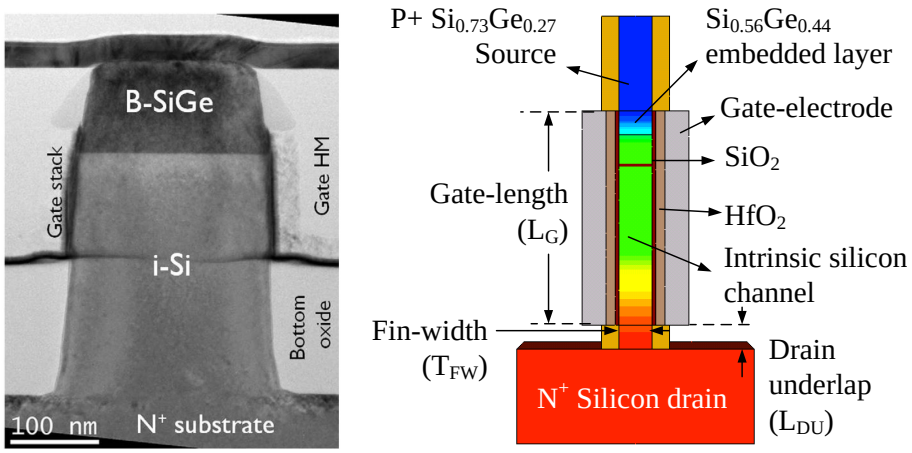
### 5.2.1 Simulations of TFET

Since TFET is an emerging-technology and its test-structures are not commercially available, technology-computer-assisted-design (TCAD)-based simulations are used for the exploration of TFET. The electrical-characteristics of TFETs of various channel-materials are extracted using commercial TCAD simulator Sentaurus Device from Synopsys [73]. A non-local BTBT model is used. A non-local BTBT model dynamically searches for the tunneling-path by following the steepest energy-band gradient at varying bias conditions. Because of the dynamic tunneling-path adaptation, a non-local BTBT model is more accurate than the conventional Schenk's [74] and Hurkx's [75] models. Shockley-Read-Hall (SRH) recombination model is also used for the realistic estimation of TFET off-current.

### 5.2.2 Calibration of TFET Simulation Parameters

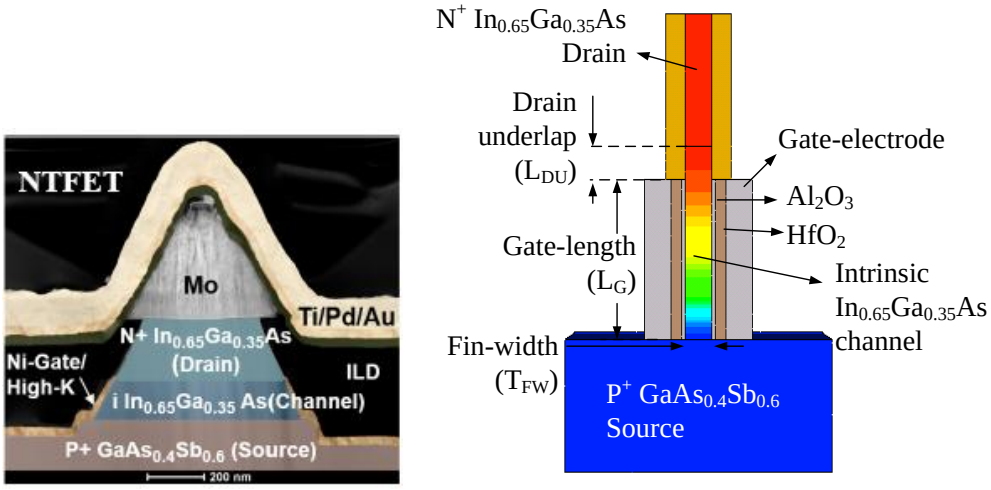
Fabrication and measurements of Si-Ge and III-V heterojunction TFETs were presented in [1, 2]. An equivalent TCAD structure is considered in Figure 31 & 32 to calibrate the simulation parameters against the measured characteristics. Various simulation parameters for Si-Ge TFET in Table 3 were obtained through the TCAD calibration of a p-Ge MOSFET and Ge-TFETs characteristics. To model TAT induced leakage, an effective $\tau_{max}$ for the Si-Ge TFET is used here in association with the Shockley-Read-Hall (SRH) recombination model [76]. Doping and material dependence on $\tau_{max}$ are ignored for the sake of simplicity

and due to a lack of the adequate measured data. Various simulation parameters for III-V TFET in Table 4 are obtained here by calibration to the measured characteristics in [2]. Similar to the Si-Ge TFET, for the III-V TFET as well, an effective $\tau_{max}$ is used.




(a) (b)

**Figure 31: Si-Ge heterojunction TFET demonstration: (a) fabricated structure in [1] and (b) equivalent TCAD structure.**




(a) (b)

**Figure 32: III-V heterojunction TFET demonstration: (a) fabricated structure in [2] and (b) equivalent TCAD structure.**

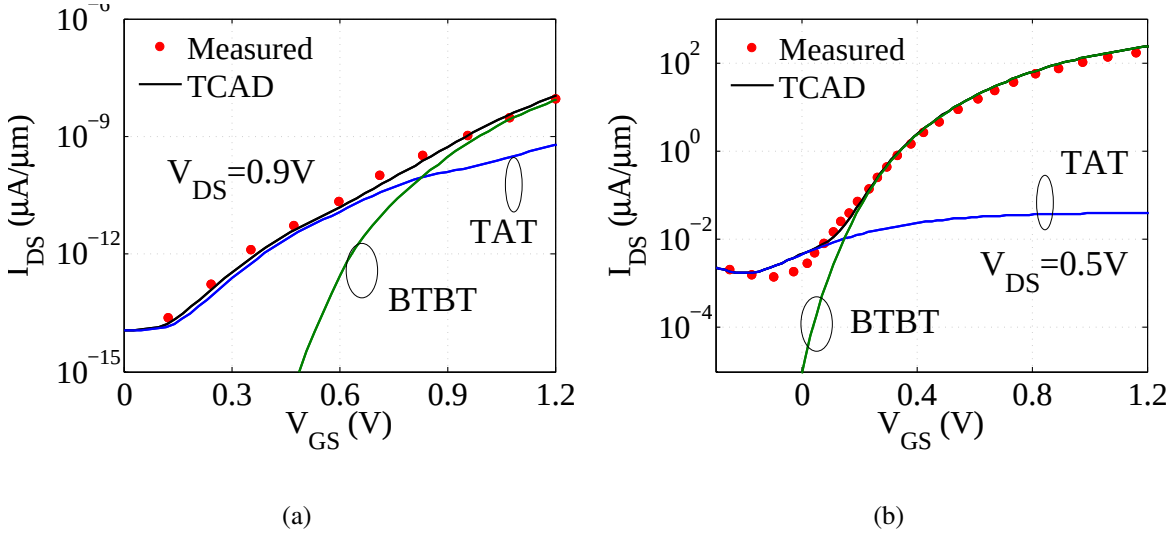
**Table 3: Si-Ge TFET simulation parameters**

| Physics | Parameter | Si | $Si_{0.73}Ge_{0.27}$ | $Si_{0.56}Ge_{0.44}$ | Ref |
|---|---|---|---|---|---|
| BTBT | $A_{path}$ (/cm$^3$/s) | $3.29\times10^{15}$ | $2.68\times10^{15}$ | $2.37\times10^{15}$ | [77] |
| | $B_{path}$ (V/cm) | $2.38\times10^{7}$ | $1.87\times10^{7}$ | $1.63\times10^{7}$ | |
| | $P_{path}$ (eV) | 0.037 | 0.037 | 0.037 | |
| Band-energy | $E_{g0}$ (eV) | Default | 1.04 | 0.92 | [78] |
| | $\chi_0$ (eV) | Default | 4.04 | 4.03 | |
| SRH recombination | $\tau_{max}$ (s) | $10^{-5}$, $3\times10^{-6}$ | $1.8\times10^{-5}$, $1.3\times10^{-5}$ | $2.3\times10^{-5}$, $1.9\times10^{-5}$ | |
| Hurks TAT | $m_t$ | 0.5, 0.5 | 0.4, 0.46 | 0.33, 0.43 | [79] |

A correlation of the measured and simulated characteristics is shown in Figure 33. In Figure 33a, the drain-current of Si-Ge TFET is dominated by the TAT component and especially for the low gate-voltages. Notably, a dominant TAT in Si-Ge TFET masks the sharper SS of BTBT. Although, TAT induced leakage is much higher in III-V TFET than Si-Ge TFET, a much improved BTBT overcomes TAT induced SS degradation.

**Table 4: III-V TFET simulation parameters**

| Physics | Parameter | $GaAs_{0.4}Sb_{0.6}$ | $In_{0.65}Ga_{0.35}As$ |
|---|---|---|---|
| BTBT | $A_{path}$ ($/cm^3/s$) | $1.68\times20^{20}$ | $1.49\times10^{20}$ |
| | $B_{path}$ (V/cm) | $7.39\times10^{6}$ | $5.38\times10^{6}$ |
| | $P_{path}$ (eV) | 0 | 0 |
| Band-energy | $E_{g0}$ (eV) | 0.87 | 0.78 |
| | $\chi_0$ (eV) | 4.19 | 4.61 |
| SRH recombination | $\tau_{max}$ (s) | $10^{-10}$, $10^{-10}$ | $10^{-10}$, $10^{-10}$ |
| Hurks TAT | $m_t$ | 0.05, 0.06 | 0.03, 0.03 |




(a) (b)

**Figure 33: Calibration of TCAD simulation parameters against hardware: (a) Si-Ge TFET and (b) III-V TFET.**

### 5.2.3 Electrical-Characteristics of TFETs of Various Channel Materials

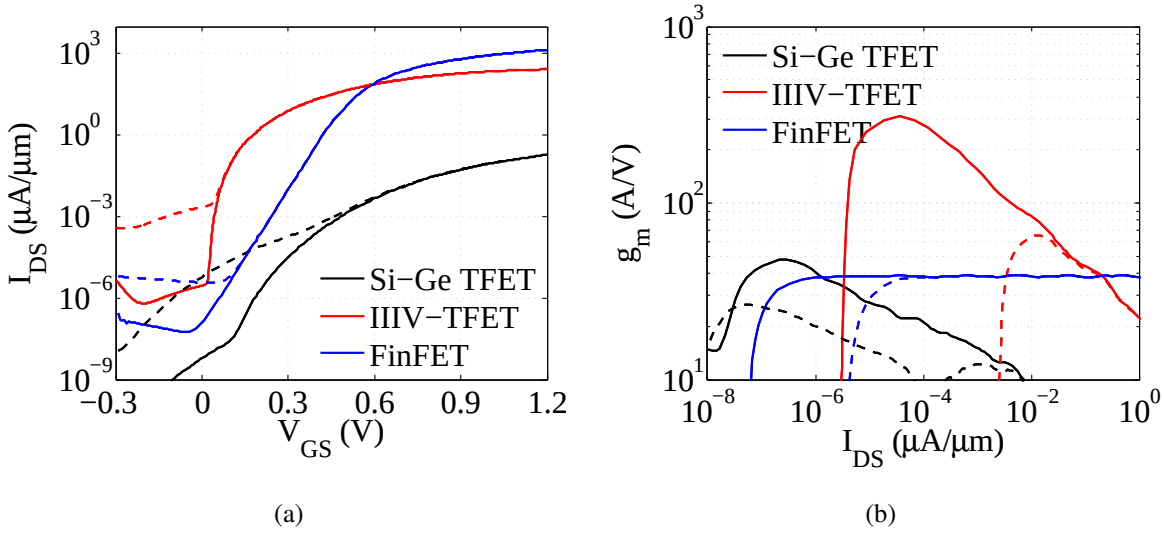
The characteristics of Si-Ge and III-V heterojunction TFET are compared considering the equivalent geometry and process adaptation shown in Table 5. The calibrated simulation parameters in Table 3 & 4 are used for the respective TFETs. A drain-underlap is adopted to suppress the ambipolar current conduction in TFETs. A steep doping-gradient (2nm/decade) is considered at the source/channel-junction in both the TFETs to enhance their on-currents. The characteristics of TFETs are also compared against an equivalent channel length (45nm) and fin-width (7nm) FinFET. The fin-height in FinFET is 25nm, and the drain (source) doping is $10^{20}$/cm$^3$. For FinFET, unified mobility model, high-$\kappa$ induced mobility degradation model, and TAT & BTBT models with the silicon parameters in Table 3 are used.

**Table 5: TFET geometrical and process specifications**

| Parameter | Si-Ge TFET | III-V TFET |
|---|---|---|
| $L_G$ (nm) | 45nm | 45nm |
| $T_{FW}$ (nm) | 7nm | 7nm |
| $L_{DU}$ (nm) | 5nm | 5nm |
| $T_{OX}$ (interfacial ox. thickness) (nm) | 0.7nm | 0.7nm |
| $T_{HK}$ (high-K ox. thickness) (nm) | 2nm | 2nm |
| Embedded layer thickness (nm) | 6nm | - |
| Source doping (/cm$^3$) | $10^{20}$ | $4\times10^{19}$ |
| Drain doping (/cm$^3$) | $5\times10^{18}$ | $4\times10^{17}$ |
| Source doping gradient (nm/decade) | 2 | 2 |
| Gate workfunction (eV) | 4.15 | 4.65 |

Drain-current ($I_{DS}$) and transconductance ($g_m$) of the various TFETs and FinFET is

shown at varying gate voltage ($V_{GS}$) in Figure 34. The characteristics with a dashed-line are for the calibrated TAT parameters as listed in Table 4. The characteristics with a solid-line are when the respective calibrated generation-recombination time-constant ($\tau_{max}$) has been scaled down by a factor of 1000. Notably, a lower $\tau_{max}$ represents a lower trap-density to suppress the TAT-induced leakage.




(a) (b)

**Figure 34: Comparison of characteristics among Si-Ge TFET, III-V-TFET, and FinFET: (a) $I_{DS}$-$V_{GS}$ and (b) $g_m$-$V_{GS}$.**

In Figure 34a, a suppressed TAT displays much lower off-current in both the TFETs and FinFET. However, containing TAT is rather more critical for Si-Ge TFET where the TAT-induced leakage masks the steep switching slope of BTBT. With a contained TAT, the Si-Ge TFET and III-V TFET show a much steeper switching-slope than in the FinFET. Bandgap of the channel and source materials controls the on-current ($I_{ON}$) and off-current ($I_{OFF}$) in TFETs. A high source and channel bandgap results in a much lower $I_{ON}$ & $I_{OFF}$ in Si-Ge TFET than in a FinFET. Meanwhile, in III-V channel TFETs, a higher $I_{ON}$ & $I_{OFF}$ is due to a low source/channel bandgap. In Figure 34b, a steeper switching-slope in both the TFETs improves their $g_m$ per bias-current (i.e., $g_m/I_{DS}$). However, the suppression of TAT is critical to obtain a higher $g_m/I_{DS}$ in both the Si-Ge TFET and III-V TFET. Notably,

with a suppressed TAT, a III-V TFET achieves several times higher $g_m/I_{DS}$ than in FinFET.

Drain current saturation in both the TFETs and FinFET is compared in Figure 35. The gate voltage in each transistor is selected such as the saturated drain current is 10 nA/$\mu$m. Notably, saturation in both the TFETs is delayed as compared to in FinFET. Si-Ge TFET incurs a higher saturation delay as compared to III-V TFET. A higher saturation voltage in TFETs will limit their output voltage swing. However, TFETs also achieve a much steady saturation than FinFET, and thereby will exhibit a higher output resistance ($r_O$) than FinFET. The drain current in TFETs is controlled by their source/channel electrostatics rather the gate/channel electrostatics as in a FinFET. Therefore, unlike a MOSFET, the drain fringing fields have a limited influence in the source/channel electrostatics of a TFET resulting in an ideal saturation and a high $r_O$.

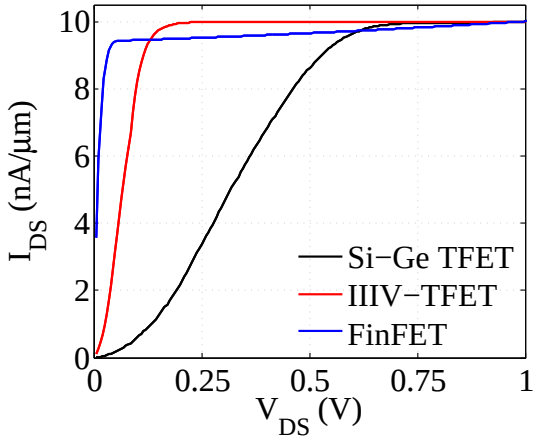



**Figure 35: Comparison of $I_{DS}$-$V_{DS}$ characteristics among Si-Ge TFET, III-V-TFET, and FinFET.**

Capacitances in various TFETs are compared in Figure 36. The gate-to-drain capacitance ($C_{GD}$) in TFET is dominant over the gate-to-source capacitance ($C_{GS}$). Meanwhile, in FinFET, $C_{GS}$ is dominant over $C_{GD}$ [80]. In TFET, a forward-biased channel/drain-junction strongly couples channel-charge with the drain-voltage resulting in a much higher $C_{GD}$. However, if $V_{GS}$ is adequately lower than $V_{DS}$, TFETs can have a lower $C_{GD}$ than FinFET

as shown in Figure 36d. Therefore, the bias conditions of TFET-based analog designs can be appropriately designed to obviate a higher $C_{GD}$ in TFET. III-V channel TFETs have a lower-capacitance than the silicon-channel TFETs because of a lower density-of-states in their channel than in silicon [9].

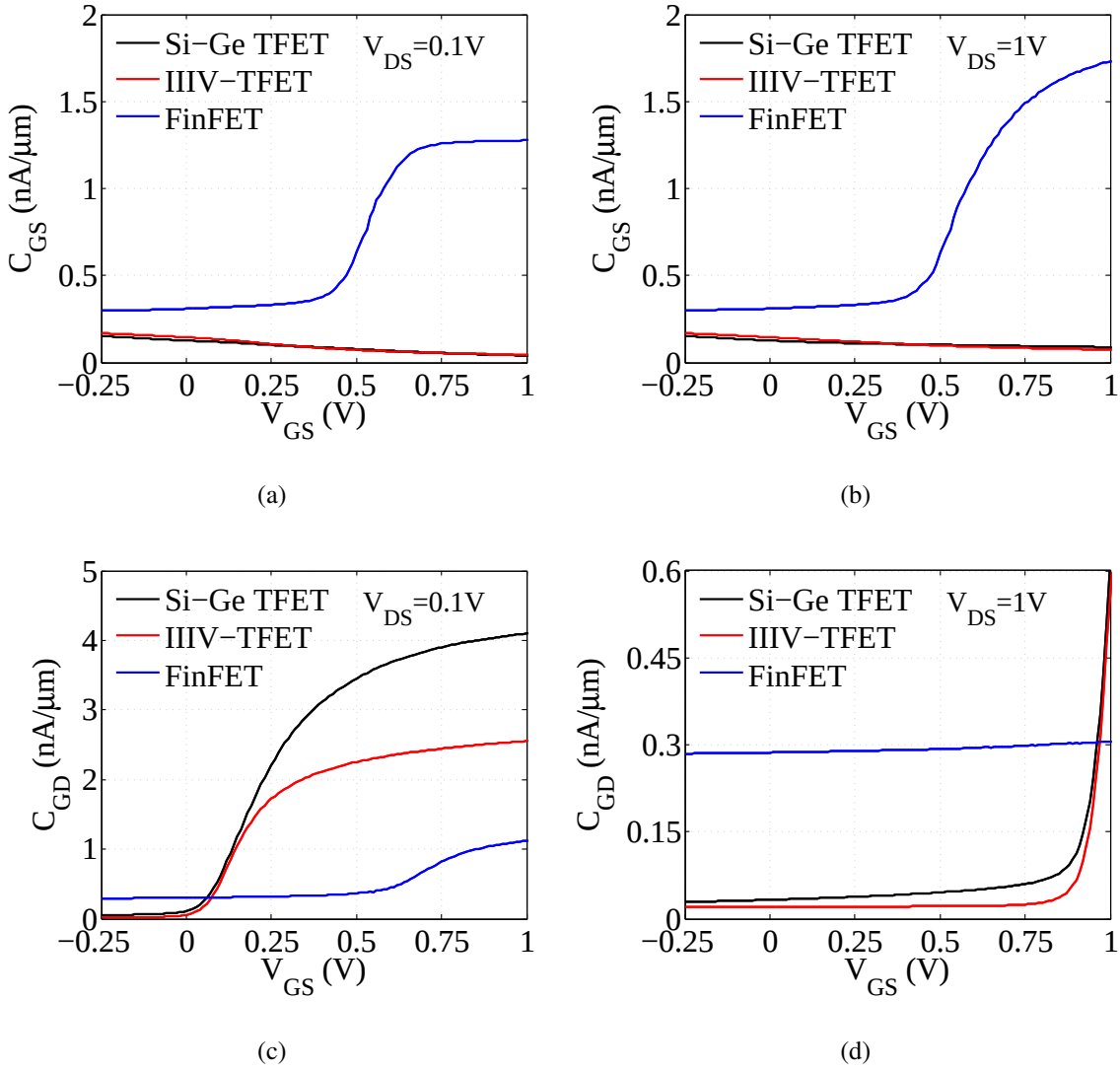



**Figure 36: Comparison of capacitance among Si-Ge TFET, III-V-TFET, and FinFET: (a) $C_{GS}$-$V_{GS}$ ($V_{DS}$ = 0.1 V), (b) $C_{GS}$-$V_{GS}$ ($V_{DS}$ = 1 V), (c) $C_{GD}$-$V_{GS}$ ($V_{DS}$ = 0.1 V), and (d) $C_{GD}$-$V_{GS}$ ($V_{DS}$ = 1 V)**

.

## 5.3 TFET-based Circuit Simulation Methodology

To explore TFET-based circuit designs, a compact model of TFET is required. Various compact-modeling works for TFET are underway [21, 81, 82]; however, a comprehensive commercial compact model for TFET is yet to appear. In lieu of a compact model, the simulation methodology shown in Figure 37 is followed for the evaluations of TFET-based circuits. In the methodology, at first electrical characteristics ($I_{DS}$, $C_{GS}$, and $C_{GD}$) of TFET are extracted using Synopsys Sentaurus TCAD simulations with bias-conditions finely varying over a wide operating-range. Next, Verilog-A based table-models are constructed by interpolating the characteristics with quadratic spline. Spice circuit-simulations are performed based-on these table models to study TFET-based circuit designs.

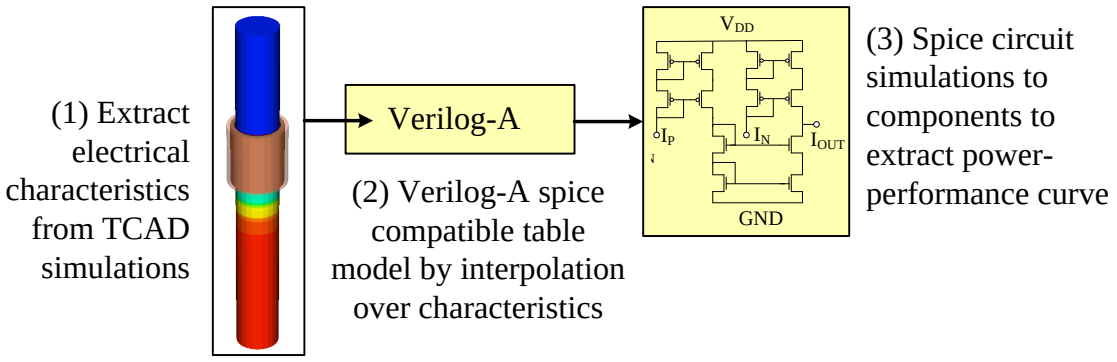
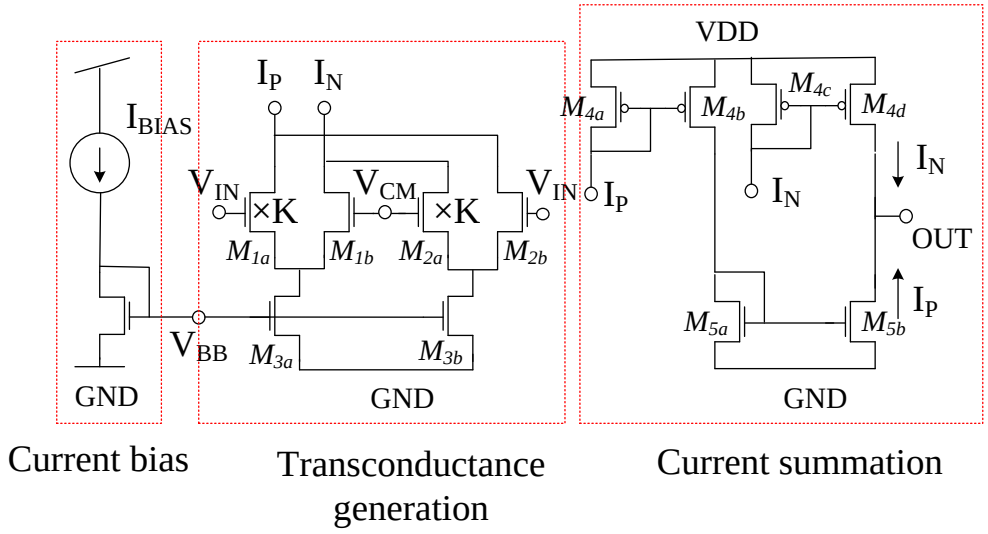



**Figure 37: Simulation methodology for TFET-based circuit designs.**

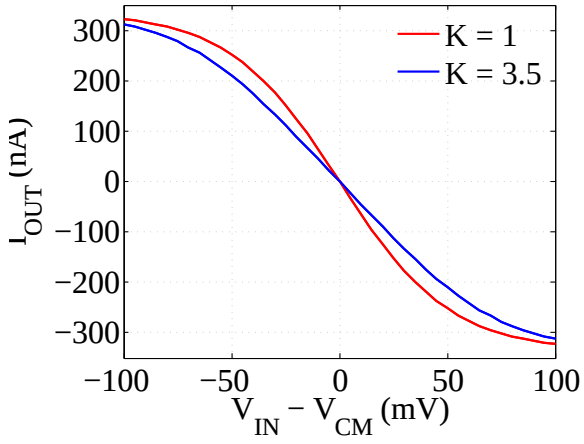
## 5.4 TFET-based Operational-Transconductance-Amplifier

Schematic of an OTA is shown in Figure 38a, where OTA consists of the transconductance generator (TG) and current summer (CS). In Figure 38b, the OTA generates an output current, $I_{OUT}$, proportional to its input voltage, $V_{IN}$. Transconductance, $GM$, of an OTA is defined as the slope of its $_{OUT}$-V$_{IN}$ characteristics. OTA transconductance (OTA-$GM$) is controlled by its bias current, $I_{BIAS}$. A cross coupled configuration at the TG stage, as shown in Figure 38a, expands the linearity of OTA. Transistor M$_{1a}$ (M$_{2a}$) is sized 'K' times higher than M$_{1b}$ (M$_{2b}$). The net transconductance (i.e. G$_{m,T}$)) in a cross-coupled OTA achieves a higher linear range, when the transconductance of each of the cross-coupled

pair compensates for the roll-off in the other. In Figure 38b, with a higher 'K', the linearity of cross-coupled OTA increases. The optimal 'K' is also dependent upon the switching slope of the transistor. Particularly, for TFET since its switching-slope depends on $V_{GS}$, the optimal 'K' is also dependent on the bias-current through the OTA.




(a)




(b)

**Figure 38: (a) Schematic of cross-coupled operational-transconductance-amplifier (OTA). (b) $I_{OUT}$-$V_{IN}$ of the OTA at varying design factor K.**

The key characteristics of OTA designed by SiGe-TFET, III-V-TFET, and FinFET are compared in Figure 39. OTAs-based on the various transistors have the same geometrical

specifications as shown in Table 6. However, the width parameter K is optimized for each OTA at varying operation power to limit distortion from linearity. A linearity range of 40 mV and -40 mV across $V_{CM}$ is considered for each OTA. The maximum distortion from linearity occurs at the limits of the linearity range, i.e., at the input voltages $V_{TP}$ and $V_{TN}$ (here, $V_{TP}$ = 40 mV & $V_{TN}$ = -40 mV). The maximum distortion from linearity is obtained by best-fitting the OTA $I_{OUT}$-$V_{IN}$ against a cubic polynomial, $f(x) = p_3 \times x^3 + p_2 \times x^2 + p_1 \times x + p_0$. The distortion from linearity is defined as $V_{TP}^2 \times |p_3|/|p_1|$. The OTA design parameter 'K' is optimized so that the distortion from linearity is < 10%.

**Table 6: TFET-OTA design specifications**

| | | | |
|---|---|---|---|
| VDD | 0.8V | $M_{1b}$ ($M_{2b}$) | 100 nm |
| $M_{3a}$ ($M_{3b}$) | 500 nm | $M_{4a}$ ($M_{4c}$) | 200 nm |
| $M_{4b}$ ($M_{5a}$) | 200 nm | MR | 3 |

In Figure 39a, $GM/P_{OTA}$ of various OTAs are compared across bias power by varying the bias current $I_{BIAS}$. The ratio of the OTA-$GM$ to its bias power, $P_{OTA}$, can be expressed as

$$\frac{GM}{P_{OTA}} = \frac{1}{2VDD} \times \sum_{i=1}^{2} \left( \frac{g_m(M_{ia})}{I_{DS}(M_{ia})} \frac{K}{K+1} + \frac{g_m(M_{ib})}{I_{DS}(M_{ib})} \frac{1}{K+1} \right) \times \frac{MR}{MR+1} \tag{17}$$

Here, $K$ is the transistor width ratio for $M_{1a}$ to $M_{1b}$ ($M_{2a}$ to $M_{2b}$), and MR is the mirror ratio of current summation stage, i.e., the ratio of width for $M_{4b}$ to $M_{4a}$ ($M_{4d}$ to $M_{4c}$).

From Equation 17, higher $g_m/I_{DS}$ of input transistors improves $GM/P_{OTA}$. Since TFETs have higher $g_m/I_{DS}$ at lower $I_{DS}$ due to steepening SS, $GM/P_{OTA}$ of TFET-OTAs increases at the lower bias power. Si-Ge-TFET-OTA at < 1 pW and III-V-TFET-OTA at < 50nW have higher $GM/P_{OTA}$ than even an ideal MOSFET-based OTA. The ideal MOSFET-OTA characteristics are extracted considering SS = 60mV/decade. In contrast, since SS of FinFET is invariant, the FinFET-OTA has a relatively invariant $GM/P_{OTA}$ across bias power.

At very low bias power, $GM/P_{OTA}$ of Si-Ge TFET and FinFET-based OTAs degrades due to short channel effects. However, Si-Ge TFET-based OTA is more scalable in power than the FinFET-based OTA. An even higher $GM/P_{OTA}$ in III-V-TFET-based OTA is constrained due to linearity limitation.

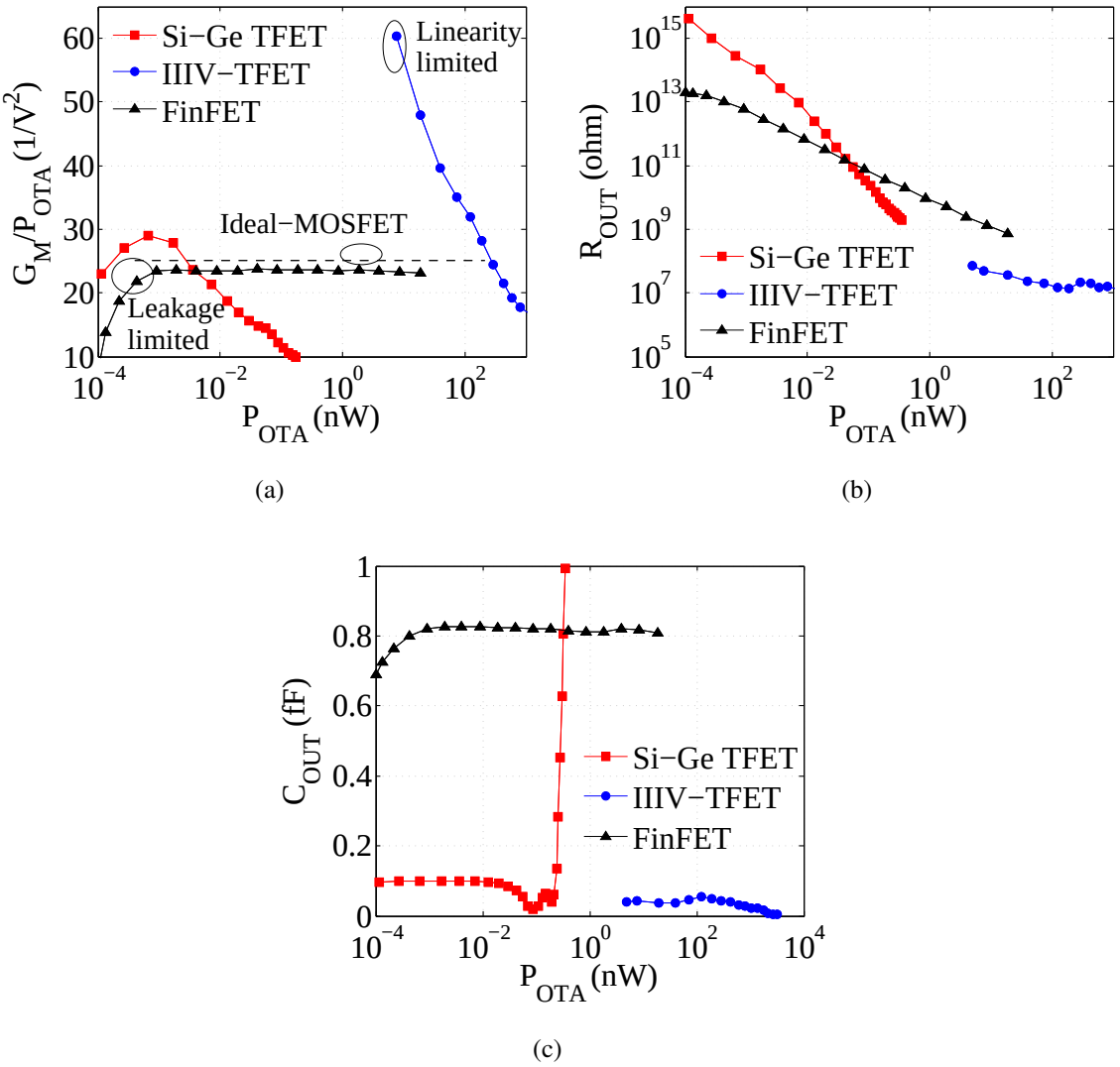



**Figure 39: Comparison of OTA-characteristics among SiGe-TFET, III-V-TFET, and FinFET across operating power: (a) transconductnace per power (GM/$P_{OTA}$), (b) output resistance ($R_{OUT}$), and (c) output capacitance ($C_{OUT}$).**

The output resistance ($R_{OUT}$) of various OTAs is compared in Figure 39b across varying

OTA bias-power. $R_{OUT}$ is defined by the small signal input resistance at the node 'OUT' when $V_{IN} = V_{CM}$. $R_{OUT}$ decreases at higher $P_{OTA}$ due to increasing bias current across OUT node (i.e., in branch $M_{4d}$-$M_{5b}$). Since Si-Ge TFET has an improved saturation, at the same very low bias-power Si-Ge-TFET-based OTA has a higher $R_{OUT}$ than FinFET-OTA. However, $R_{OUT}$ in Si-Ge-TFET-OTA falls at a sharper rate with increasing $P_{OTA}$ due to a delayed saturation in Si-Ge TFET.

The output capacitance ($C_{OUT}$) of various OTAs is compared in Figure 39c across varying OTA bias-power. $C_{OUT}$ is defined by the small signal input capacitance at the node 'OUT' when $V_{IN} = V_{CM}$. $C_{OUT}$ is mainly determined by $C_{GD}$ of $M_{4d}$ & $M_{5b}$ in the OTA schematic. At a low bias power operation, $C_{OUT}$ of TFET-OTAs is low since TFET-$C_{GD}$ is suppressed by the gate-drain under-lap. However, at a higher bias power operation, as the channel of $M_{4d}$ & $M_{5b}$ in SiGe-TFET-OTA inverts, the $C_{OUT}$ of Si-Ge TFET-OTA suddenly rises and exceeds from the FinFET-TFET-OTA. III-V-TFET-OTA has a lower $C_{OUT}$ than Si-Ge-TFET-OTA due to a lower density of states in III-V TFET materials than silicon.

## 5.5 Conclusions

This chapter has investigated the energy-efficiency of TFET-based analog designs. Energy-efficiency of MOSFET-based analog designs is limited due to a limited switching slope in MOSFET. TFET due to its subthermal switching slope facilitates higher energy-efficiency in analog designs than MOSFET. TFET-based OTA was studied in this chapter. A Verilog-A table-based simulation methodology integrating TCAD-based device simulations and HSPICE-based circuit simulations was used to study the TFET-based OTA. TFET-based OTA achieves a higher $GM/P_{OTA}$ than even an ideal MOSFET-based OTA provided TAT induced leakage current in TFET can be contained. III-V-TFET-OTA achieves much higher $GM/P_{OTA}$ than Si-Ge-TFET-based OTA due to a much steeper switching slope in III-V-TFET. TFET displays an improved saturation and thereby a higher output resistance in TFET-based OTA than in MOSFET-based OTA. However, TFET-based OTA also incurs a

limited output swing due to a delayed saturation in TFET, and a higher output capacitance due to a higher $C_{GD}$ in TFET. The bias conditions in TFET-based OTA, however, can be designed to limit the channel inversion induced higher $C_{GD}$, and a lower output capacitance than even MOSFET-based OTA can be achieved under these optimized bias conditions.

# CHAPTER 6

# TFET CELLULAR NEURAL NETWORK-BASED IMAGE PROCESSING

There is a growing demand for ultra-low-power image processing platforms. Applications such as bionic eye [83], smart camera pixel [84] require moderate to high image throughput within a stringent power budget. Neural network based image processing is shown to be much more adept than the conventional digital platforms [85]. In particular, cellular neural networks (CNNs) [33, 35] have attracted attention due to its VLSI favorable architecture. This chapter particularly focuses on the power reduction in CNN synapse through post-CMOS device TFET. Note that in CNN even in its simplest nearest neighborhood 2D architecture [35], each neuron requires 16 synapses to propagate/receive the CNN dynamics, and the synapse count grows rapidly for more complex (e.g. higher neighborhood radius [33]) architectures. Therefore, power-reduction in TFET synapse is critical, and TFET-based OTA is investigated for this goal. Various contributions discussed in this work were published in [86, 71].

## 6.1 Cellular Neural Network Cell Design Constraints

The CNN cell design was shown in Figure 5. A CNN cell integrator accumulates output current from the various OTAs, while the output current of these OTAs is controlled by the input and output voltages of the neighboring cells. The input resistance ($R_{int}$) and input capacitance ($C_{int}$) of the integrator is given as

$$R_{int} = R/A_{int} \tag{18a}$$

$$C_{int} = A_{int} \times C. \tag{18b}$$

Here, R & C are the resistance and capacitance of the CNN cell, respectively. And, $A_{int}$ is the closed-loop gain of the integrator amplifier. Thereofre, in order to reliably sink

current from the neighboring cell OTAs to the integrator,

$$A_{int}/R >> \Sigma_n 1/R_{OTA,n} \tag{19a}$$

$$A_{int} \times C >> \Sigma_n C_{OTA,n}. \tag{19b}$$

Here, $R_{OTA,n}$ & $C_{OTA,n}$ are the output resistance and capacitance of the neighboring OTA. Therefore, the integrator op-amp gain is determined so that Equation 19 is satisfied. Furthermore, for the stability of the CNN, the necessary and sufficient condition based on the feedback coefficient matrix A shown in [35]

$$A(2,2) > 1/R. \tag{20}$$

Therefore, the CNN cell resistance apart from Equation 19 should also follow Equation 20.

## 6.2 Cellular Neural Network Cell Power Scaling

### 6.2.1 CNN Cell Power Scaling Approach

A power-scaling approach for CNN is discussed without altering the equilibrium or steady state outputs of the CNN, and satisfying Equations 19 & 20. In Equation 4b, with increasing R, scaling down the coefficients $A_{kl,ij}$, $B_{kl,ij}$, and $I_{ij}$ inversely proportional to R (i.e., $\propto$ 1/R), results into the same equilibrium states (i.e. $V(x_{ij})$ at which d$V(x_{ij})$/dt = 0). Hence, with the increasing R, the cell bias current $I_{ij}$ and OTA-GM to generate the coefficients $A_{kl,ij}$ and $B_{kl,ij}$ can be reduced. Therefore, with the increasing R, along with the cell bias power, the OTA bias power can also be reduced. Although, with the increasing time constant of CNN cell (i.e. R×C) the settling time increases, the equilibrium states of the network are retained across operating power.

### 6.2.2 CNN Cell Power Scaling Simulation Results

CNN cell power scaling is studied for noise-filtering applications considering design constraints and power-scaling approach described above. To satisfy Equation 20, R = 1.1/A(2,2), where A(2,2) is the feedback transconductance for noise-filtering template. To satisfy Equation 19, a close-loop gain ($A_{int}$) of 25 is targeted in the integrator-opamp. A closed

loop-gain $A_{int} >> 1$ is also necessary for the functionality of integrator. $A_{int}$= 25 is also sufficient to satisfy Equation 19. To satisfy Equation 19, the CNN cell capacitance C is selected as $\frac{10}{A_{int}} \times \Sigma C_{OTA,i}$ where $C_{OTA,i}$ is the output capacitance of the peripheral OTA to the integrator.

In Figure 40, the synapse-biasing-power of a CNN-cell is compared for FinFET, SiGe-TFET, and III-V-TFET-based implementations across varying time-constants. Because of a superior GM/$P_{OTA}$ and lower $C_{OTA}$, SiGe-TFET-based synapse requires lower biasing power than FinFET at the same CNN time-constant configuration. However, as the SiGe-TFET synapse biasing power increases, $C_{OTA}$ in the TFET-OTA increases due to a channel inversion in its output transistors [Figure 39c], and thereby the realized time-constant in the TFET-OTA degrades. IIIV-TFET-based OTA, meanwhile, is less scalable in power due to a higher OFF current in IIIV-TFET. However, IIIV-TFET-based OTA realizes a higher performance operation (i.e., lower CNN time-constant) at a much lower power than FinFET-OTA.

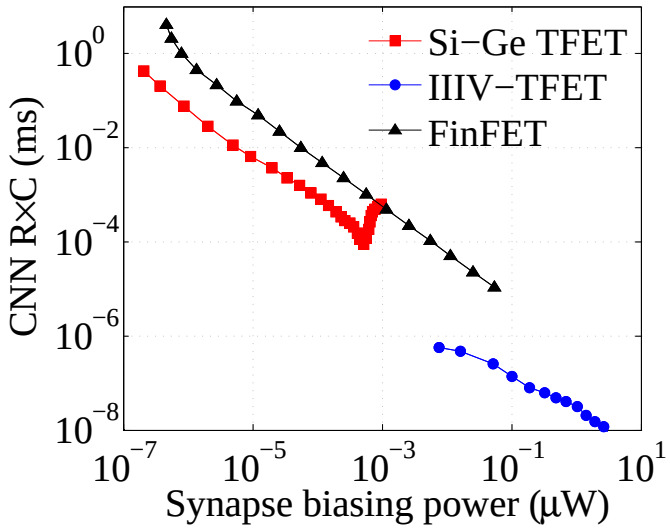



**Figure 40: CNN synapse power comparison among SiGe TFET, IIIV TFET, and FinFET-based implementations at the varying CNN cell time-constant.**

Utilizing the above CNN power-scaling approach, with the increasing R, as the output impedance for the integrator increases, the biasing power of the integrator op-amp can also

be reduced. A seven-transistor OPAMP configuration is used for integrator implementation as shown in Figure 41a. In Figure 41b, the bias power of SiGe-TFET, IIIV-TFET, and FinFET-based OPAMP is compared at varying R×C load (as obtained from the CNN cell power scaling). At a very low power operation, OFF current dominate the biasing power in FinFET-based integrator. Therefore, at a high R×C, SiGe-TFET-based integrator requires lower biasing power than FinFET-based integrator. However, at a lower time-constant operation, SiGe TFET-based integrator requires higher power than FinFET-based integrator due to a degraded $g_m/I_{DS}$ in SiGe-TFET with higher current bias. Similar to a IIIV-TFET-based synapse, the power-scaling of a IIIV-TFET-based integrator is also limited due to a higher OFF current in IIIV-TFET.

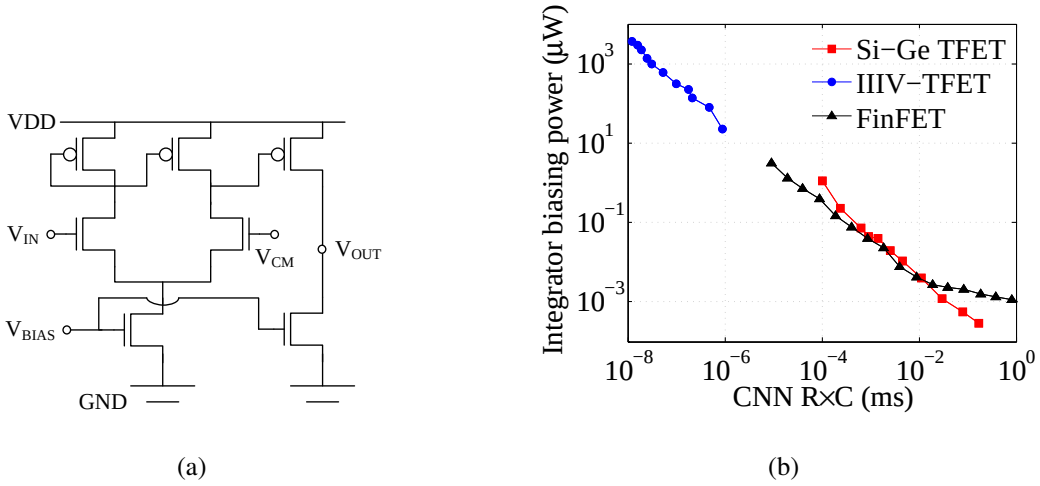



**Figure 41: (a) Schematic of a seven transistor OPAMP used in the integrator. (b) For a closed-loop gain of 25, the integrator power at varying R×C time-constant.**

## 6.3 Simulation Methodology for CNN-based Image Processing

CNN-based image processing is evaluated by a cohesive simulation methodology utilizing TCAD simulations of TFET and FinFET, HSPICE simulations of TFET & FinFET-based circuits, and functional simulations on MATLAB of network dynamics. The functional simulations of Equation 4b determine the network convergence time or the throughput for

a given image processing step. Circuit simulations of CNN-cell are used for the power-estimation of the network. Utilizing the power-GM trace of TFET and FinFET-based OTA in Figure 40, the CNN synapse bias power of CNN is estimated. Utilizing the power-R×C trace of TFET and FinFET-based integrator in Figure 41b, the CNN neuron power is estimated. Power contributions through saturating function generator are ignored. The saturating function generator is implemented as a unity gain amplifier [36], and it should follow the same power trend as an integrator. The dynamic power dissipation is evaluated by integrating voltage evolution on the CNN-cell capacitance until the network equilibrium. Therefore, an integrated simulation framework utilizing functional simulations of MATLAB, circuit simulations of HSPICE, and TCAD device simulations estimates the throughtput efficiency (i.e., throughput per power) of CNN-based image processing for TFET and FinFET-based implementation. The net CNN power is given by

$$P_{CNN} = \sum_{ij} \left( P_{dyn}(V_{x,ij}) + V_{DD} \times I_{ij} + \sum_{kl \in S_{ij}} P_{OTA}(A_{kl}) + \sum_{kl \in S_{ij}} P_{OTA}(B_{kl}) + P_{OPAMP}(R, C) \right) \quad (21)$$

Here, i,j are the CNN cell index and $S_{ij}$ is the cell of neighboring cells for $C_{ij}$. $P_{dyn}(V_{xij})$ is the dynamic power dissipated across cell capacitance C in the CNN cell $C_{ij}$. $P_{OPAMP}$ is the integrator bias power.

## 6.4 Comparison of TFET and FinFET CNN-based Image Processing

TFET and FinFET CNN-based image processing is compared for a 'fixed-array approach' and an 'energy-optimal approach'. These approaches are defined as below:

### 6.4.1 Fixed-Array Approach

Throughput efficiency of TFET and FinFET CNN-based image processing is first compared in Figure 42 by following a fixed-array approach. A 66×66 CNN array is utilized to process a 514×514 image as shown in Figure 6 for noise-filtering and edge-detection. A multiplexing procedure from [36] is utilized to process a larger resolution image than

the CNN array. The array bias power is varied to modulate the network convergence time-constant while following the earlier described CNN power-scaling approach. Throughput of the network increases with the increasing operating power of CNN and the reducing time-constant. Throughput-efficiency (i.e., no. of image processing operations per second per power) is compared for TFET and FinFET-based CNN for the above image processing steps. The energy-optimal approach is also noteworthy from the perspective that it overcomes the degrading performance of a TFET-CNN cell at a higher power operation.

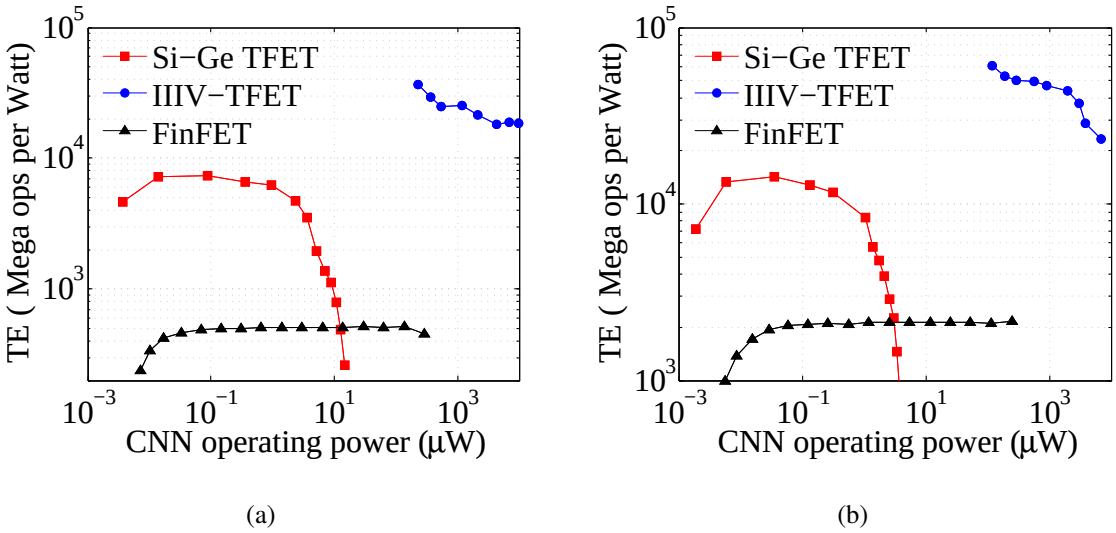



**Figure 42: Throughput efficiency of SiGe-TFET, IIIV-TFET, and FinFET-based CNN at varying operating power under fixed array approach for (a) edge-detection, and (b) noise-filtering.**

At a lower power operation, SiGe-TFET-based CNN is much more energy-efficient than a FinFET-based CNN. SiGe-TFET utilizes its higher $g_m/I_{DS}$ and lower $C_{GD}$ under lower current bias to enhance energy-efficiency of SiGe-TFET-based CNN. However, at a higher power operation, the throughput-efficiency (TE) of SiGe-TFET-based CNN significantly reduces. At a higher bias power, SiGe-TFET has degraded $g_m/I_{DS}$ and higher $C_{GD}$ than FinFET which leads to a degraded TE in SiGe-TFET-based CNN. IIIV-TFET-based CNN is much more energy-efficient and suitable for high performance operations. However, power-scalability of IIIV-TFET-based CNN is constrained due to a higher OFF-current in

IIIV-TFET.

### 6.4.2 Energy-optimal Approach

Although, the multiplexing procedure from [36] enables processing a larger resolution image than the CNN array size, the multiplexing introduces inefficiency due to the overlap between the successive steps, and thus, processing more effective pixels than image. Considering an 'energy-optimal' approach, where the CNN array size is rather determined by the maximum number of minimum powered cells within the array power specifications, the TE of TFET and FinFET-based CNN is compared. The minimum power of FinFET synapse is chosen as 0.8 pW, since beyond this power the efficiency of synapse reduces [Figure 39a]. Similarly, considering power-efficiency limitations, the minimum power of SiGe-TFET synapse is chosen as 0.6 pW. The minimum power of IIIV-TFET synapse is 5 nW. The TE of various TFET and FinFET-based CNN designs is compared in Figure 43.

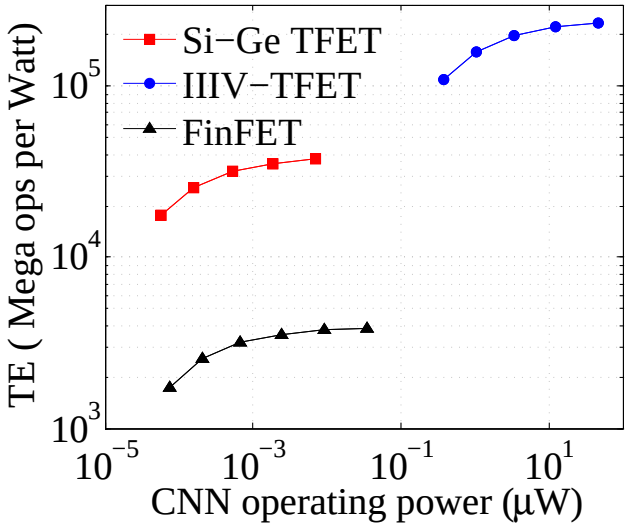



**Figure 43: Throughput efficiency of SiGe-TFET, IIIV-TFET, and FinFET-based CNN at varying operating power under energy-optimal approach for edge-detection.**

By minimizing the number of processed image pixels and by operating the CNN cell at an energy optimal point, the above energy-optimal approach improves the TE from the previous fixed-array approach.

## 6.5 Conclusions

This chapter has shown that TFETs, due to its steeper SS, are attractive for ultralow-power image processing through CNN. Although, the low on-current of Si/Ge-TFET limits its utility in digital designs; with cellular parallelism, a CNN can still extract higher performance from Si/Ge-TFET. While the current research focus in TFET is toward improving its on-current for digital applications; this study underscores the importance of the TFET design with steeper SS and lower off-current for cellular neural network-based computing.

# CHAPTER 7

# TFET CELLULAR NEURAL NETWORK-BASED ASSOCIATIVE MEMORY

This chapter studies the application of TFET in designing a low power and robust cellular neural network (CNN)-based associative memory (AM). A TFET-OTA is utilized as a programmable synaptic weight multiplier for CNN. The ultralow-power of TFET-OTA enables a higher connectivity network even at a lower power, and thereby improves the memory capacity and input pattern noise tolerance of CNN-AM for low power applications. Since a higher connectivity network is critical for CNN-AM, the chapter particularly focuses on SiGe-TFET-based CNN-AM due to a much greater power-scalability of SiGe-TFET-based OTA than a IIIV-TFET-based OTA. Various contributions discussed in this chapter were published in [87].

## 7.1 Motivation for CNN Associative Memory with TFET

Applications of AM have been investigated in solving problems, such as character/face recognition, pattern classification, database search, and understanding/replicating cerebral activities [88, 89, 90]. While performance requirement for the problems such as recognition and classification are moderate, an ultralow power of AM can enable solving these complex problems in a low power platform, such as in a mobile system-on-a-chip. Ultimately, an ultralow-power of AM can also enable an ambitious goal of a very large scale AM computing, such as in a mammalian brain with $10^{10}$ neurons (biological computing elements) and $10^{14}$ synapses (biological interconnects), at sustainable operating power. Hence, a critical requirement for an AM computing platform is to minimize its power while meeting the throughput and performance demands. While operating at lower power, AM should be able to correctly identify the correlation (referred to as a successful recall), even under noise in the input pattern and maximize the total number of stored associations (defined as

memory-capacity).

Cellular neural network (CNN) has been investigated for AM applications [39]. The algorithmic analysis of CNN-AM has shown that increasing the cell-to-cell connectivity, i.e., having more connections per cell, improves successful recall and memory-capacity [91] as shown in Figure 44.

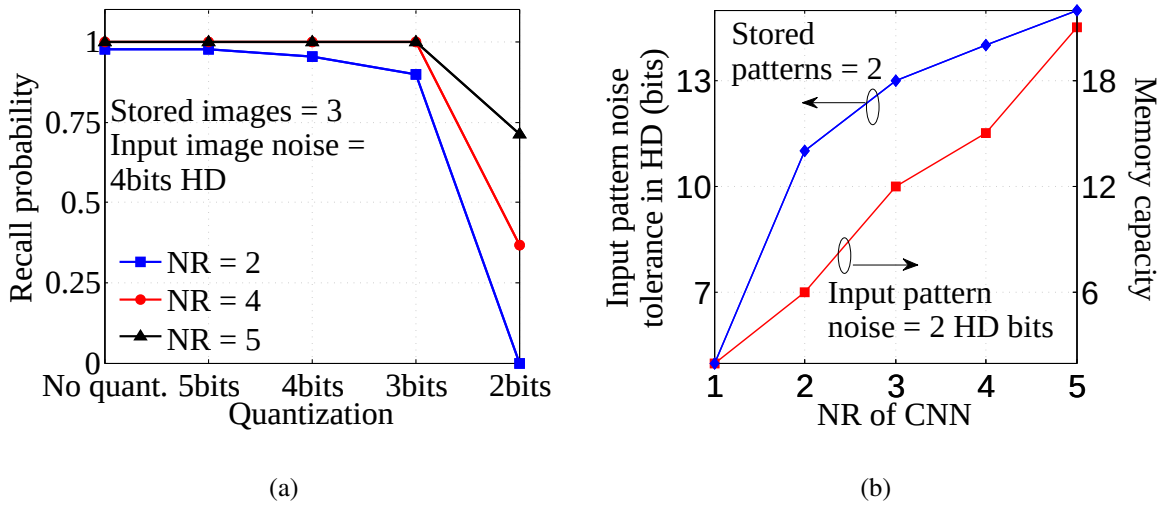



**Figure 44: (a) Recall probability at varying degree of quantization and for varying NR CNN-AM. (b) Memory capacity and input noise tolerance (in HD bits), at varying NR. Results are for 11×11 CNN-AM.**

A higher cell-to-cell connectivity implies a nonlinear increase in the number of synaptic-weight multipliers per cell. Therefore, the ability to design ultralow-power synaptic-weight multiplier becomes critical for low power AM. In Chapter 5, the benefits of TFET for ultralow power analog design were presented. This chapter utilizes a TFET-OTA as a synaptic-weight multiplier for CNN, and explores a TFET-based low power, robust, and high performance CNN-AM platform.

## 7.2 TFET-based CNN-AM Simulation Methodology

An integrated methodology connecting device simulations using TCAD, circuit simulations using SPICE, and functional simulations using MATLAB for CNN-AM study is presented in Figure 45. In CNN-AM, for a given cell resistance R, the Hebbian learning algorithm [92] determines the synaptic-weights (OTA-GMs) $A_{ij,kl}$, $B_{ij,kl}$, and bias, $I_{ij}$, to store correlation between the desired patterns. For these parameters, MATLAB solves functional simulations of CNN-AM to estimate performance (recall speed, input pattern noise tolerance, and memory-capacity). Recall speed is defined by the duration when various state voltages of the CNN array saturate to 95% of their equilibrium values as shown in Figure 46. CNN-AM power is scaled and estimated using the methodology discussed in Chapter 6.

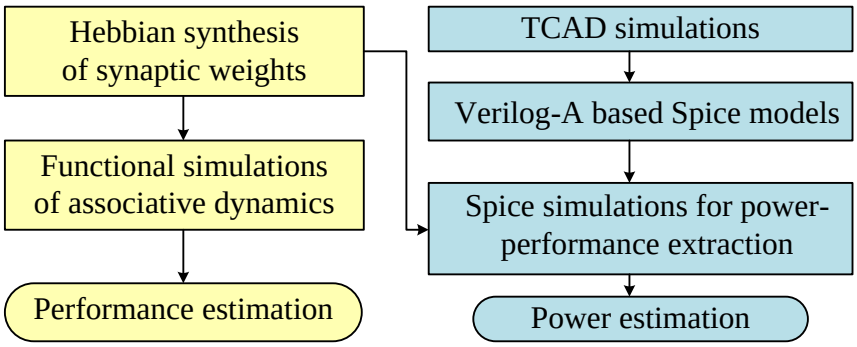



**Figure 45: Cohesive simulation methodology, integrating TCAD, SPICE, and functional simulations to extract CNN-AM characteristics at different technologies, TFET and FinFET.**

## 7.3 TFET-based CNN-AM Simulation Results

### 7.3.1 CNN-AM at NR of One

The power-performance characteristics of Si-Ge TFET, III-V TFET, and FinFET-based CNN-AM is compared with NR of one. A five bit least square signed quantization of the synaptic-weights is considered due to the complications of accurate analog weight storage. With the quantized synaptic-weights, the circuit schematic of Figure 47 will enable digital storage/distribution of space varying analog synaptic-weights. Here, the tail transistors of

OTA, $M_{3a}$, and $M_{3b}$, are implemented with quantized widths. And, a number of instances of $M_{3a}$ and $M_{3b}$ are selected to implement the tail biasing of OTA depending on the localized storage of synaptic-weights.

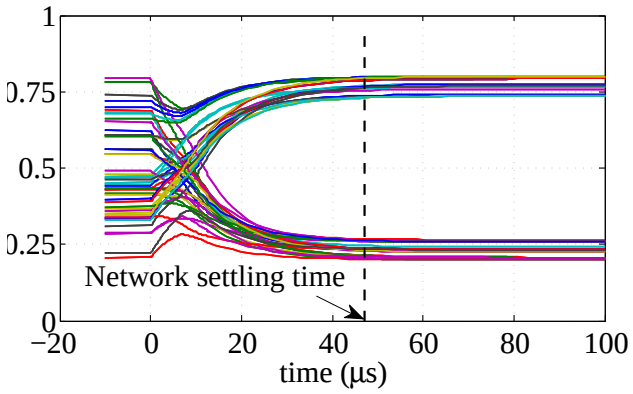



**Figure 46: Transient evolution of various cell state voltages, and the network settling time.**

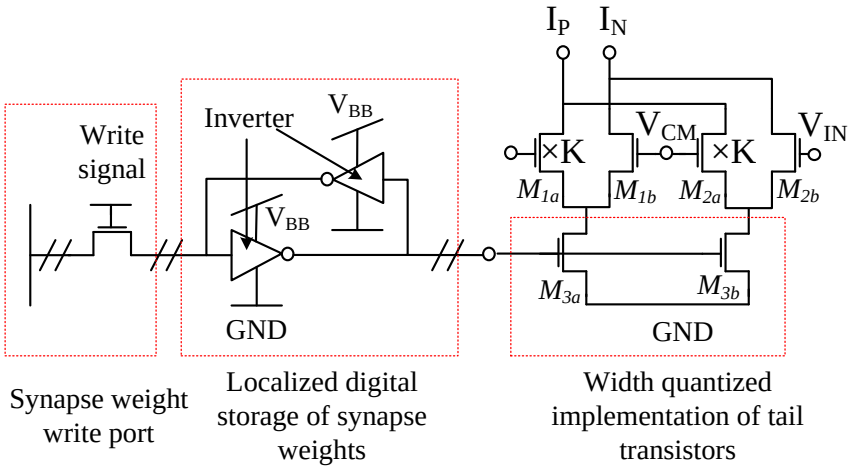
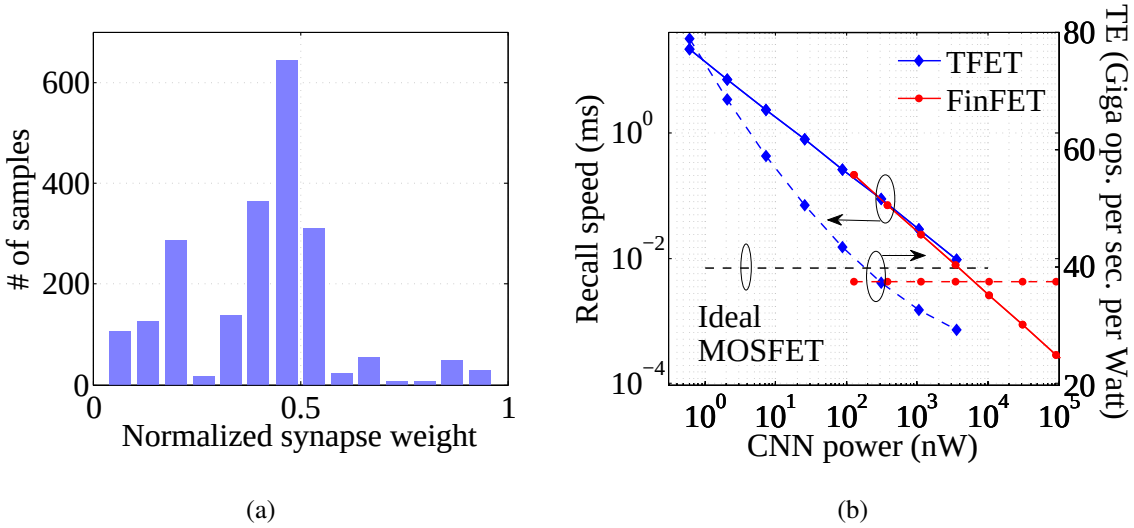



**Figure 47: Circuit scheme to locally store and implement quantized synapse weights.**

Simulation methodology as described in the previous section is utilized. Best-fit power-performance characteristics of TFET and FinFET-OTA are utilized. In Figure 48a, a distribution of quantized synaptic weights is shown. The power-scaling approach scales each of these coefficients inversely proportional to R, and therefore, the bias power to realize corresponding GMs also reduces. Power-scaling of OTAs is limited to the region where leakage

currents do not overwhelm transconductance current (i.e., GM/$P_{OTA}$ doesn't exhibit a leakage induced roll-off). Hence, the minimum power and minimum realizable GM in TFET- and FinFET-OTA is limited; this, in turn, limits the power-scaling through the earlier approach and the minimum operational power for CNN. In Figure 48b, the TFET-CNN-AM operates down to ~nW power due to the ultralow power scalability of TFET-OTA, and ~ms recall time at this power will still be useful for recognition and classification applications in low power systems.




(a) (b)

**Figure 48: (a) Distribution of synapse weights. (b) Recall speed and throughput-efficiency for TFET- and FinFET-CNN-AM across CNN power.**

The net OTA bias power distribution at varying synaptic weights for TFET- and FinFET-CNN-AM is shown in Figure 49a for low performance operation (recall speed = 1 ms). A higher GM/$P_{OTA}$ of TFET-OTA at low power attributes to lower bias power across synaptic-weights in TFET-CNN-AM than FinFET-CNN-AM. However, at a higher power since GM/$P_{OTA}$ of TFET-OTA is lower, the TFET-CNN-AM has higher power than FinFET-CNN-AM for such high performance operation [Figure 49b, recall speed = 10 $\mu$s]. Therefore, a TFET-CNN-AM can only achieve higher throughput efficiency (TE) than FinFET-CNN-AM as long as the performance constraints are low.

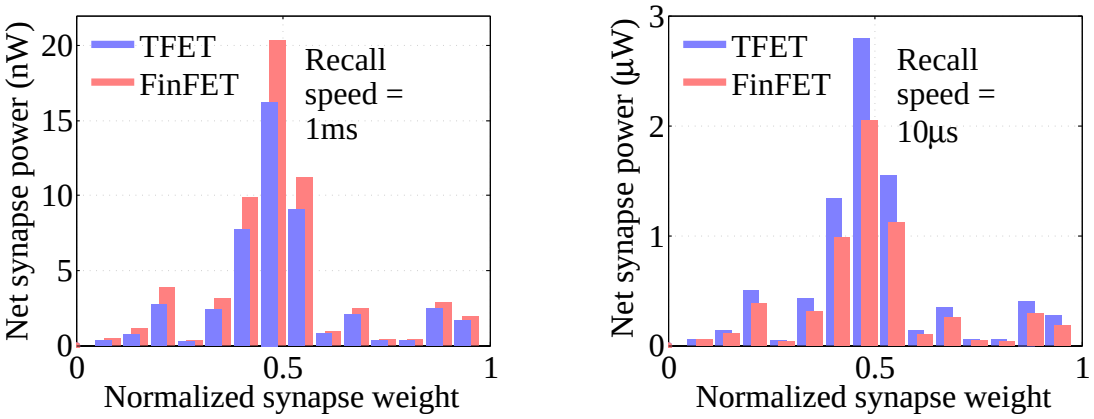



**Figure 49: The synapse power distribution and comparison between TFET -and FinFET-CNN-AM for (a) low and (b) high performance application.**

In Figure 48b, the TE of TFET-CNN-AM and FinFET-CNN-AM is compared across recall speed and CNN power. Since, GM/$P_{OTA}$ of TFET-OTA improves at lower bias current, the TE of TFET-CNN-AM also improves at lower power. Below 200 nW, TFET-CNN-AM has better TE than even an ideal MOSFET (i.e., with switching slope = 60 mV/decade)-based CNN-AM. However, note that under subthreshold operation, due to a constant switching slope of FinFET, GM/$P_{OTA}$ of FinFET-OTA is invariant across power, hence, the TE of FinFET-CNN-AM is also invariant across power under subthreshold operation of FinFET-OTA.

### 7.3.2 Improving TFET-CNN-AM by High NR Design

Throughput efficiency of a TFET-CNN-AM can be improved by a higher NR design. In a higher NR design, there are more OTAs per cell. Thus, at a given total power for CNN, in higher NR design, the power allocated for each of the OTA is lesser, and the realized OTA-GM is lower. Meanwhile, GM/$P_{OTA}$ of TFET-OTA increases at a lower power. Therefore, at the same total power, in a higher NR design, TFET-OTAs operate at a more energy-efficient point. Due to this unique characteristic of TFET, and with higher OTA count in a

higher NR CNN, a higher net transconductance ($\Sigma$OTA-GM$_n$) can be realized for a higher NR TFET-CNN cell at the similar power [Figure 50]. In Figure 51, TE characteristics of TFET-CNN-AM are demonstrated for higher NR designs. Although, a higher NR TFET-CNN-AM design is less scalable in power, it achieves better throughput efficiency at the similar power. Also, note that these throughput efficiency benefits of high NR design are unique to TFET-CNN-AM. Due to a constant GM/P$_{OTA}$ of FinFET-OTA, the throughput efficiency in FinFET-CNN-AM is limited, and a higher neighborhood-radius (NR = 2) does not improve the throughput efficiency.

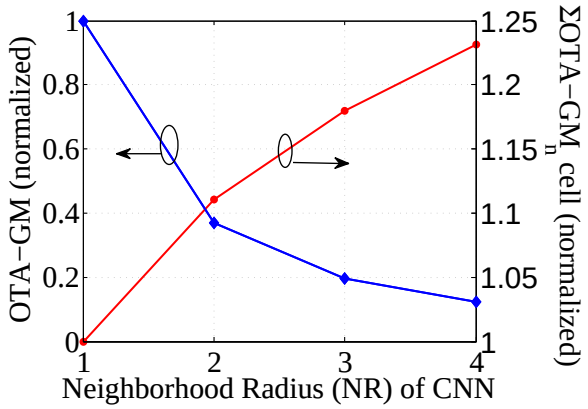



**Figure 50: The OTA transconductance and net CNN cell transconductance across NR for iso-powered CNN designs.**

Therefore, an optimal approach to exploit the higher GM/P$_{OTA}$ of TFET-OTA is by operating the TFET-CNN-AM at the maximum NR given the CNN power scaling limitations as shown in Figure 51. Furthermore, a higher NR operation can also achieve the higher algorithmic quality [Figure 44]. In Figure 52, utilizing a maximum NR TFET- and FinFET-CNN-AM, the algorithmic quality is shown at varying power. For Figure 52a, the test-case of Figure 7 is considered. In Figure 52b, for the memory-capacity test, apart from the patterns of Figure 7, varying count of additional random binary patterns are considered for synthesis and recall. TFET, by enabling a higher NR CNN-AM even at a lower power, enables a higher algorithmic quality. At 200 nW, while TFET based NR = 5 CNN-AM

enables hamming distance noise tolerance of 15 bits and memory-capacity of 21 patterns, the FinFET based NR = 1 CNN-AM only achieves 5 bits and two patterns respectively.

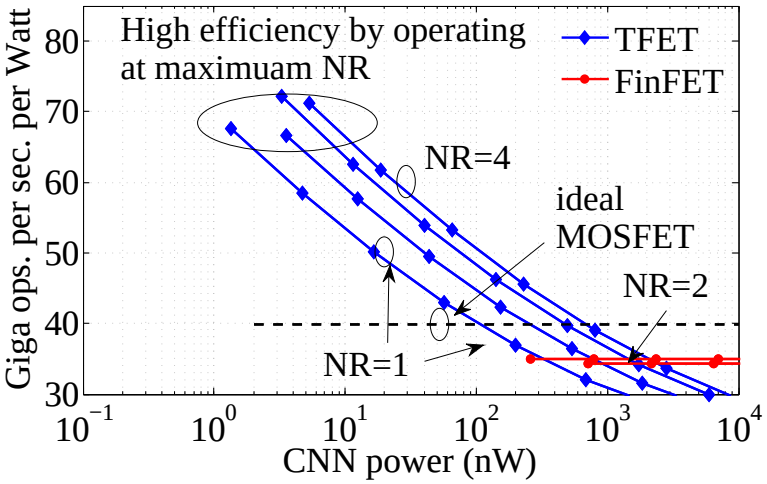
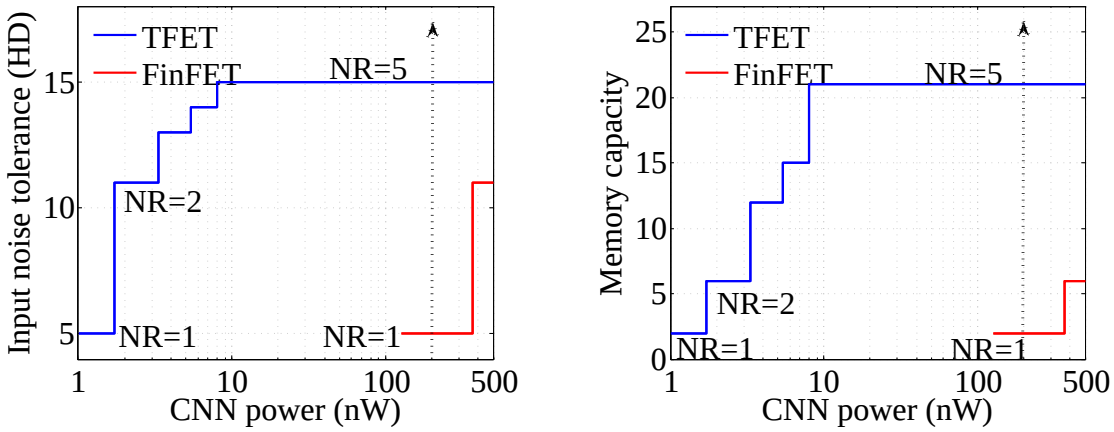



**Figure 51: Throughput efficiency (TE) of TFET- and FinFET-CNN-AM at varying NR.**




(a) (b)

**Figure 52: For maximum NR operation between TFET and FinFET CNN-AM at varying power: (a) input pattern noise tolerance, and (b) memory capacity.**

## 7.4 CNN-AM under Process Variations

The impact of transistor variability on AM quality is studied through functional simulations of CNN-AM. Transistor variability induces non-idealities in OTA, as offset voltage (VO) and transconductance error ($GM_E$ rather than the designed $GM_O$). Considering a Normal distribution on VO and a log-Normal distribution on $GM_E/GM_O$, variability is introduced to the synaptic-weights, $A_{ij,kl}$ and $B_{ij,kl}$, in the functional simulations of CNN-AM. A log-Normal distribution for $GM_E/GM_O$ is used due to the exponential sensitivity of current to gate voltage in both subthreshold FinFET and TFET. Higher transistor variability induces a greater variability in VO and GM. In Figure 53, a higher NR CNN shows significant resiliency against increasing $\sigma(GM_E/GM_O)$ and $\sigma$(VO). A higher NR CNN is more robust in AM operation to begin with [Figure 44], and the design softly fails with the higher unreliability of OTAs. Hence, a higher NR CNN design, apart from greater throughput efficiency and algorithmic quality, will also be resilient to process defects.

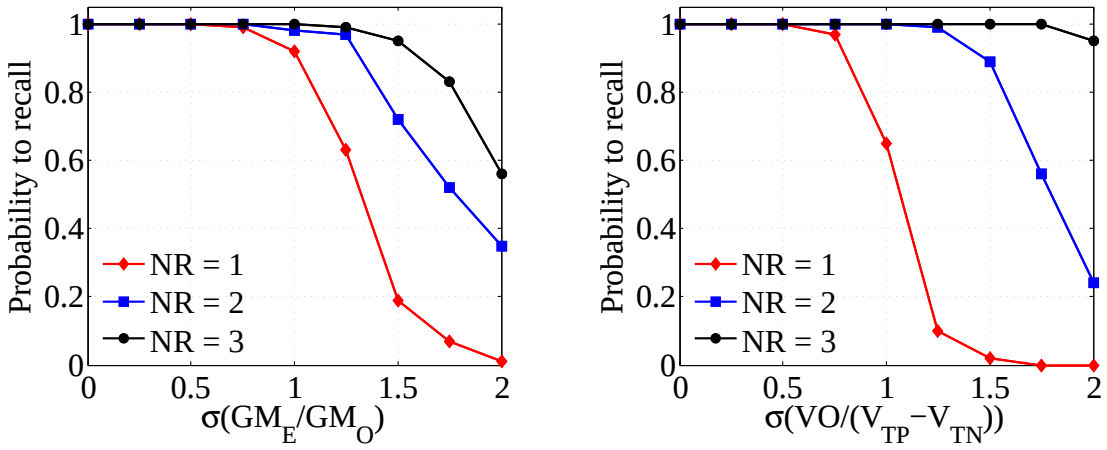



**Figure 53: Probability to recall across NR architectures with increasing variability in: (a) GM and (b) VO (normalized by saturation limit of $f_{sat}$).**

## 7.5 Implementation Complexity in a Higher NR TFET-CNN

The TFET enables implementation of a higher NR CNN even at a low power, and thus, provides opportunities for a low power AM design with higher algorithmic quality, process variation resiliency, and higher TE. However, a higher NR implementation of CNN will also increase the complexity and area of implementation. Figure 54 demonstrates algorithmic and TE of TFET-CNN-AM with increasing number of synaptic interconnections at varying NR. Note that HD noise tolerance and TE increment with higher number of synaptic interconnections begins to saturate, meanwhile, the network complexity increases proportionally to the count of interconnections. Hence, due to the area and complexity constraints in an AM design, the optimal NR will be limited.

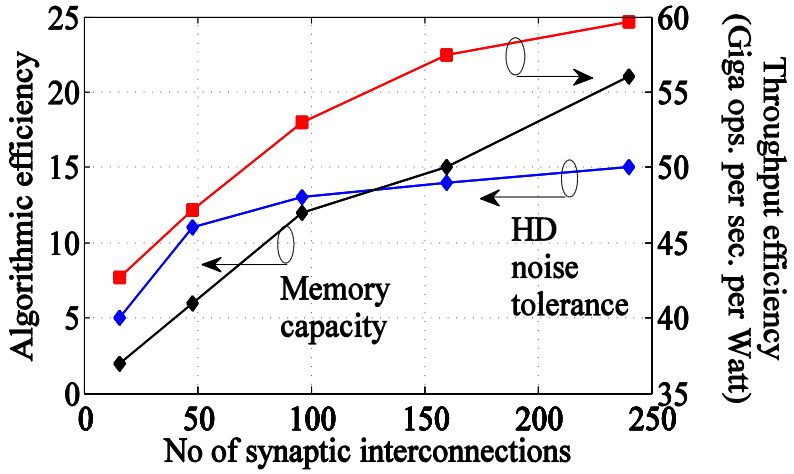



**Figure 54: Saturating HD noise tolerance and throughput efficiency at increasing number of synaptic interconnections.**

Various technological innovations such as a vertical and low footprint implementation of TFET [93] can reduce the network complexity and area requirement, and enhance the optimal NR of design. Simultaneously, novel circuit techniques can be explored to mitigate the impact of higher interconnect capacitance at higher NR. For example note that, in Figure 5, a higher interconnect capacitance does not significantly affect the functionality of CNN, since the output node of OTAs is regulated through integrator, and thus, the effect of

parasitic capacitance is suppressed.

## 7.6 TFET-CNN for Image Processing vs. Associative Memory

Notably, CNN-AM exploits the TFET characteristics differently, and perhaps more effectively, than in simple CNN based image processing applications. In Chapter 6, for TFET-CNN based image processing applications, TFETs are exploited to increase the number of nodes in the network for a given power (by reducing the OTA power) and keeping the connectivity (NR = 1) constant. More nodes lead to higher parallelism, and hence, higher throughput efficiency by exploiting parallelism in CNN based image processing. A lower off-current of transistors becomes more critical in expanding the network size and exploiting the parallelism benefits.

On the other hand, a TFET-CNN-AM is significantly benefited both from the low off-current as well as higher $g_m/I_{DS}$ of TFET as presented in this chapter. It is shown that lower off current, and hence, the low-power of TFET-OTAs are better exploited by increasing the NR in CNN-AM (under power constraints). From an algorithmic perspective, higher NR improves algorithmic quality (noise tolerance and memory-capacity). Interestingly, from an algorithmic perspective throughput efficiency is expected to be independent of NR; as observed in case of the FinFET-CNN-AM (Figure 51). However, as explained in Figure 51, the higher $g_m/I_{DS}$ of TFET changes the cell dynamics at higher NR leading to higher throughput efficiency in TFET-CNN-AM.

Furthermore, unlike space invariant templates in CNN based image processing, synapse weights in AM depend on the pattern itself, are space variant, and can vary widely in magnitude (e.g. 15× variation between the largest and smallest synapse weight in Figure 48a). Therefore, the interactions between the CNN-AM power and the variable switching slope in TFET becomes significantly more important to consider.

## 7.7 Conclusions

This chapter has presented the potential of SiGe-TFET in designing efficient and robust CNN-AM. A lower OFF-current of SiGe-TFET enables lower power operation of CNN synaptic-weight multiplier enabling a higher NR CNN for a given power. A higher NR CNN-AM improves algorithmic quality of AM. In addition, higher NR also improves TE of TFET-CNN-AM at a constant power, thanks to the steeper switching slope (higher gm per unit bias power) of TFET. Increasing performance benefits along with the increasing tolerance against process variability at higher NR indicates building higher NR CNN-AM as the suitable approach to build large scale higher performance and robust AM implementation. However, increasing implementation area along with interconnect complexity can ultimately limit the NR of implementation. The application of TFET in CNN-AM also reveals more involved device-algorithm interactions, than what observed in TFET-CNN-based image processing [Chapter 6]. Increasing NR, as performed here for CNN-AM, more effectively exploits unique TFET and AM characteristics for quality, noise tolerance, speed, power, and TE. Future work needs to consider the design challenges of higher NR CNN-AM including area and interconnect. The vertical orientation of SiGe TFET alleviates the area constraints, feature size scaling of the nanowire will be important. The local connectivity in CNN help to partially mitigate the challenge of global interconnects. The analog communication also reduces the density; however, the requirements of higher local interconnect density for higher NR design will still be a challenge.

# CHAPTER 8

# NON-CONVENTIONAL III-V HETEROJUNCTION TFET FOR AREA AND ENERGY-EFFICIENT NEUROMORPHIC ASSOCIATIVE PROCESSING

A non-conventional gate/source-overlapped Heterojunction-Tunnel-FET (SO-HTFET) with Gaussian-$I_{DS}$-$V_{GS}$ characteristics is shown in this chapter. The SO-HTFET designs a single-transistor distance-computing-cell (DCC) for associative-processing (AP) with reinforced-learning. The application of SO-HTFET-based AP to face-recognition demonstrates a higher-accuracy, 250× lower-power, and 100× higher DCC-density than a digital-CMOS-based Boolean-AP. Various contributions discussed in this chapter were published in [94, 95].

## 8.1 Distance-Computation-based Neuromorphic Architecture for Associative-Processing

Associative-processing (AP) analyzes similarity between two patterns. The basic operation in AP is to compute the difference between two inputs, performed using distance-computing-cells (DCCs). A distance-computation-based neuromorphic AP-architecture, as shown in Figure 55, couples the outputs of DCCs to make the high-level similarity decisions. A digital CMOS-based Boolean-AP requires complex logic for DCC making the design power/area-inefficient [96]. This has led to non-Boolean-AP studies. Shibata et al. showed CMOS based analog-DCC, and studied its application to AP [97]. Recently, Parihar et al. studied a coupled-relaxation-oscillator based AP platform which exploits metal-to-insulator transition in $VO_2$ [98, 96]. Requirements of a high-performance and energy-efficient AP are listed in Figure 56. For scalability of AP, DCCs must be of high-density and should require simpler inter-cell-coupling. The AP-architecture should require simpler energy-efficient peripherals for higher parallelism in AP. DCCs should also possess an on-line learning capability to exhibit AP-plasticity. With the AP-plasticity, the stored database

can be adapted in run-time depending on the input queries. A high-density, low-power, and high-performance non-Boolean-AP using HTFET to achieve the above requirements is discussed in this chapter.

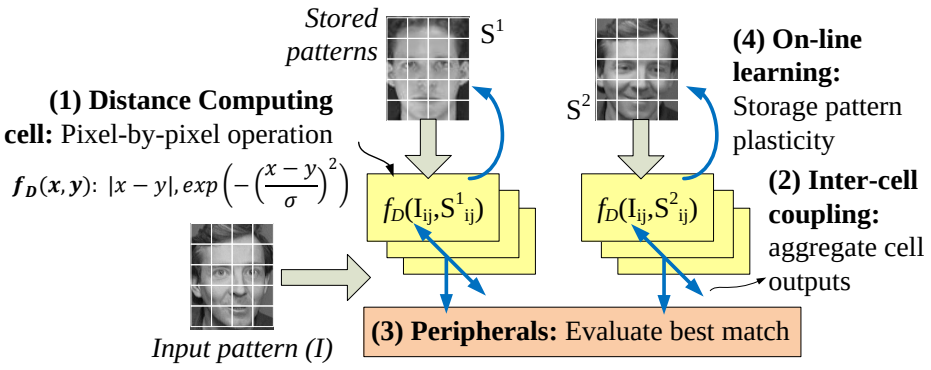



**Figure 55: Computing elements and architecture of a distance-based neuromorphic-associative-processing.**

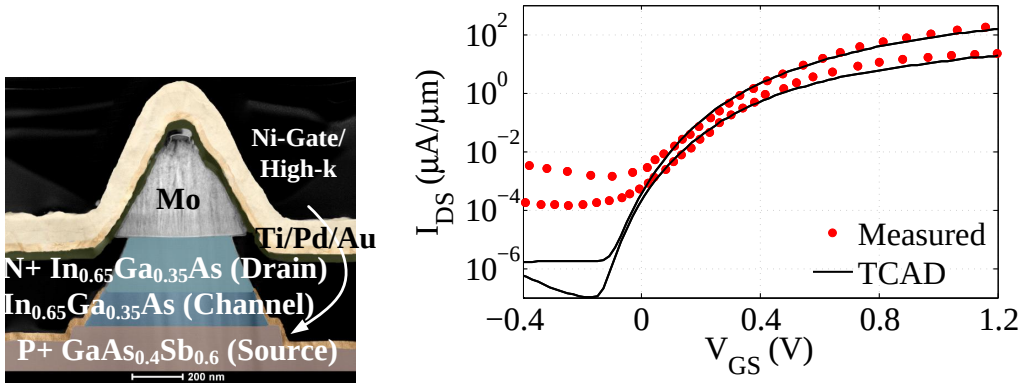
| Goals of AP | Requirements | Contributions in this work |
|---|---|---|
| **Scalability** | High density distance computing cell | Distance computing in a single floating-gate gate/source overlapped HTFET |
| | Low complexity inter-cell coupling | Steady state current based cell output, aggregation by parallel computing cells |
| **Parallelism** | Low complexity, scalable peripherals | HTFET based higher sensitivity Winner-take-all |
| **Plasticity** | On-line learning | On-line learning scheme in FG SO-HTFET array. Runtime adaptation of stored patterns. |
| **Improved accuracy** | Device-AP co-design | Device geometry, process, material optimization. |

**Figure 56: Requirements of an energy-efficient associative-processing, and contributions in this chapter.**

## 8.2 Gate/Source-overlapped Heterojunction-TFET as a Single-Transistor Distance-Computing-Cell

III-V complimentary HTFETs were shown in [2]. The n-HTFET fabricated in [2] is modified to design a single-transistor DCC by introducing gate/source-overlap (i.e., an SO-HTFET). A TCAD model calibrated against the experimental data of n-HTFET is adapted

to evaluate the SO-HTFET-based DCC [Figure 57]. In Figure 57b, a higher off-current in the measured data is because of the interface-traps. However, simulations in this chapter mainly use the on-current of HTFET, therefore the inaccuracies in off-current prediction can be ignored. The SO-HTFET schematic is shown in Figure 58. TCAD simulates a non-local BTBT with the parameters listed in Table 4.




(a) (b)

**Figure 57: TCAD calibration: (a) a fabricated $In_{0.65}Ga_{0.35}As/GaAs_{0.4}Sb_{0.6}$ n-HTFET in [2], and (b) calibration to the measured data.**

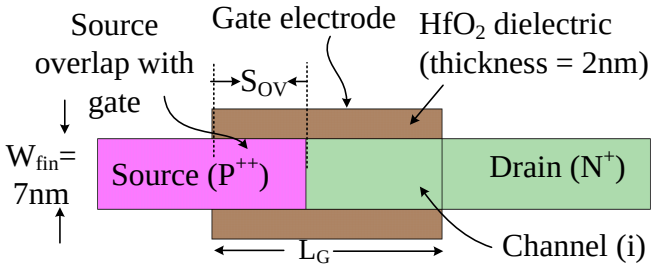



**Figure 58: Gate/Source-overlapped HTFET (SO-HTFET).**

Energy band diagrams in SO-HTFET at varying $V_{GS}$ ($V_{DS}$ = 0.2V) are shown in Figure 59. The conduction-band (CB) and valence-band (VB) energies are shown along the inter-section directly underneath the gate-electrode. Because of the gate-overlap, a positive-$V_{GS}$ in SO-HTFET induces a depletion in source near the gate-interface [compare band-profiles

without-overlap and with-overlap in Figure 59]. The depletion extends the BTBT-path-length and suppresses BTBT in the region. At much higher $V_{GS}$, the complete fin is under depletion; hetero-interface-BTBT is suppressed, and only homojunction-BTBT occurs in the source. The hole-generation profile across channel is shown in Figure 59 to further illustrate the preceding observation.

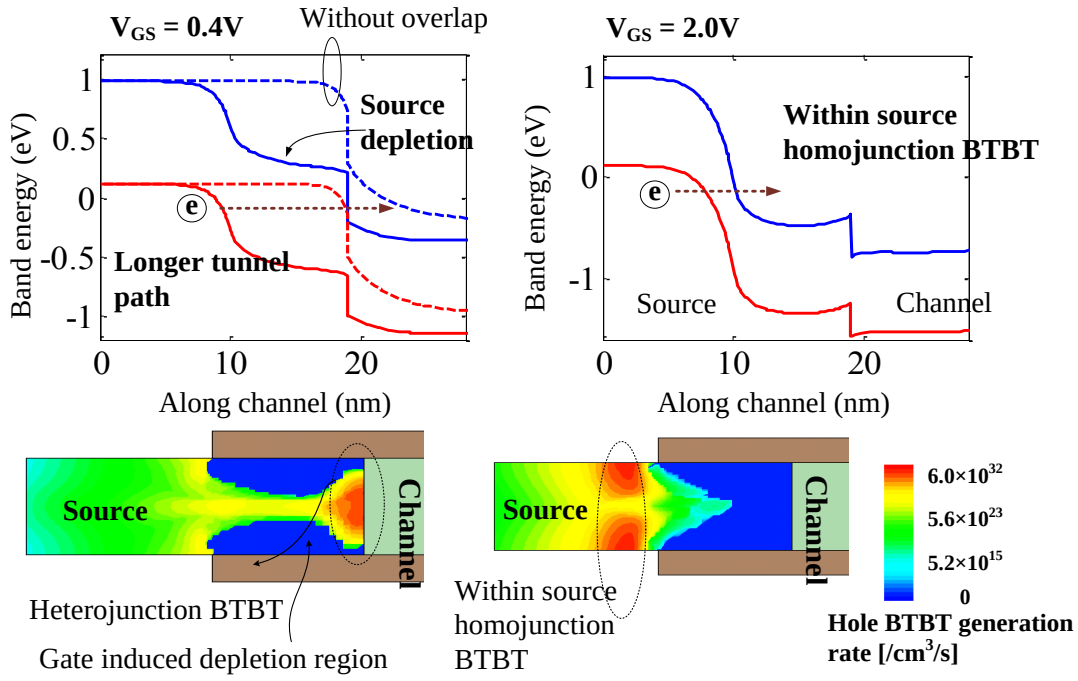



**Figure 59: Energy-band diagram and hole-generation profile in SO-HTFET.**

Contrary to a typical HTFET, because of $V_{GS}$ induced BTBT transition from the hetero-junction to homojunction mode, the SO-HTFET shows Gaussian-shaped $I_{DS}$-$V_{GS}$ [Figure 60], where $I_{DS}$ follows the difference (distance) between $V_{GS}$ and $V_{peak}$ (i.e., $V_{GS}$ at the peak $I_{DS}$). Hence, SO-HTFET achieves a 'single-transistor' DCC.

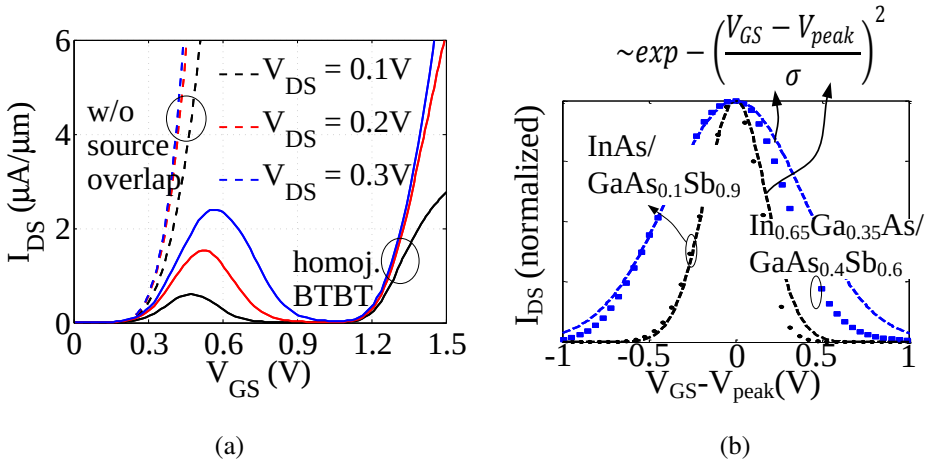



**Figure 60: Gaussian $I_{DS}$-$V_{GS}$ in SO-HTFET: (a) comparison of $I_{DS}$-$V_{GS}$ of SO-HTFET and a typical HTFET, and (b) comparison of FG-SO-HTFET $I_{DS}$-$V_{GS}$ with the standard-Gaussian characteristics.**

The geometry, doping, and material of SO-HTFET modulate the peak and variance ($\sigma$) of its Gaussian-$I_{DS}$-$V_{GS}$ [Figure 61]. The Gaussian-$I_{DS}$-$V_{GS}$ shape in SO-HTFET disappears at the wider fin-widths; since the gate-induced depletion doesn't mask the complete fin-width [Figure 61a]. At the reduced source-overlap-length ($S_{OV}$), heterojunction-BTBT is not fully suppressed at higher $V_{GS}$ [Figure 61b]. A lower source-bandgap (e.g. $GaAs_{0.1}Sb_{0.9}$) and/or channel bandgap (e.g. InAs) material in SO-HTFET enhances its peak current [Figure 61c], and reduce its Gaussian-variance [Figure 60b]. In Figure 61c, the lower-bandgap SO-HTFET-(IV) has 20× higher peak-current than the higher-bandgap SO-HTFET-(I). A reduced source-doping also suppresses homojunction-BTBT induced $I_{DS}$ rise, resulting in the Gaussian-$I_{DS}$-$V_{GS}$ shape over a wider $V_{GS}$ [Figure 61c]. In Figure 61c, the source-doping of SO-HTFET-(IV) is $1\times10^{19}/cm^3$. A sharper Gaussian-shape (low-variance/higher-peak current) is desirable for AP; hence, a thin-fin InAs/$GaAs_{0.1}Sb_{0.9}$ SO-HTFET is considered in this chapter.

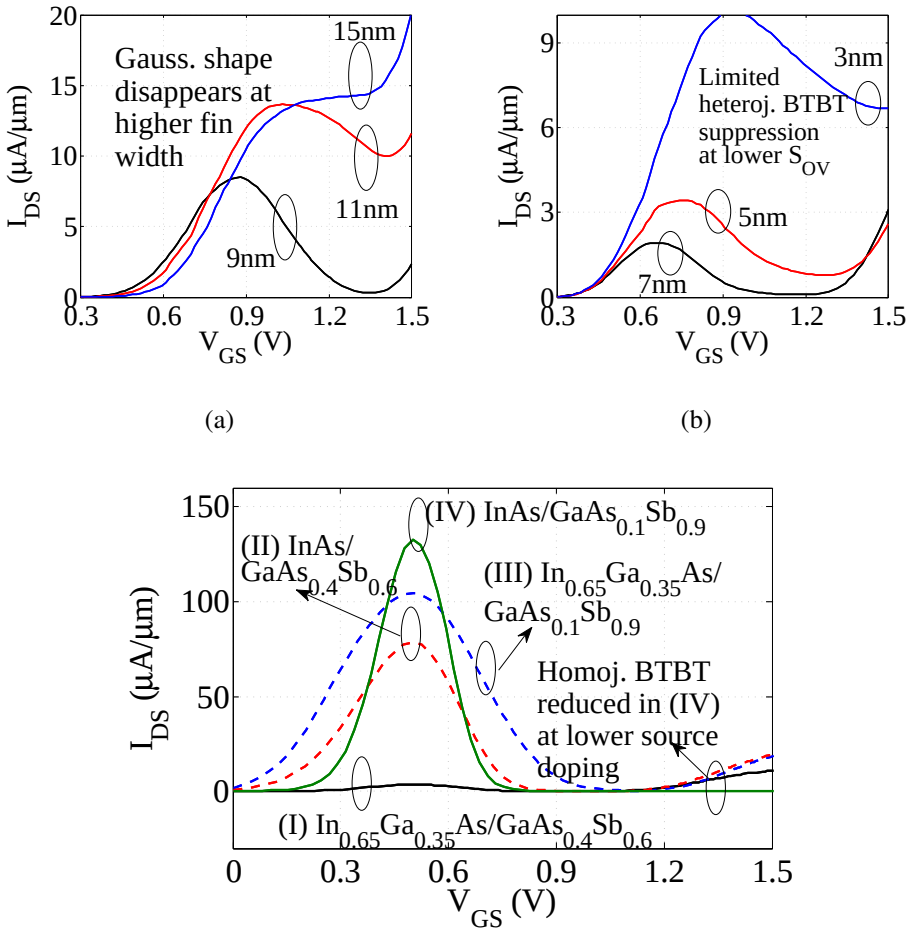



**Figure 61: Controlling Gaussian-$I_{DS}$-$V_{GS}$ of SO-HTFET: (a) at varying fin-width ($W_{FIN}$), (b) at varying gate/source-overlap length ($S_{OV}$), and (c) at varying channel/source material.**

## 8.3 Programming of SO-HTFET Distance-Computing-Cell

$V_{peak}$ in a SO-HTFET-DCC is programmed by using a floating-gate (FG) structure [Figure 62]. FG-SO-HTFET schematic is shown in Figure 62a. Thicker dielectric layers are used in FG-SO-HTFET for charge-retention in the floating-gate. $I_{DS}$-$V_{GS}$ of FG-SO-HTFET are programmed by injecting charge to the floating-gate. At sufficiently high gate-voltage, electric-field across SO-HTFET tunneling-oxide is high enough to induce charge-tunneling from the channel/source-region to the floating-gate [Figure 62b]. Injected charges are

blocked by the thicker and higher dielectric-constant blocking-oxide layer, and are trapped in the floating-gate. Note that a higher dielectric-constant in $HfO_2$ than $Al_2O_3$ induces a greater electric-field across the tunneling-oxide layer than the blocking-oxide layer, and facilitates charge-injection and trapping in the floating-gate. TCAD simulations with a direct-tunneling model are utilized to study the charge-injection to the floating-gate of FG-SO-HTFET. Direct-tunneling model [73] is suitable for both trapezoidal and triangular-tunneling barrier. Used effective-hole and electron tunneling-masses for direct-tunneling are listed in Table 7. For electrons in the source and channel-regions, Γ-valley density-of-states effective-mass is used as the tunneling-mass. For holes in the source and channel-regions, light-hole density-of-states effective-mass is used as the tunneling-mass.

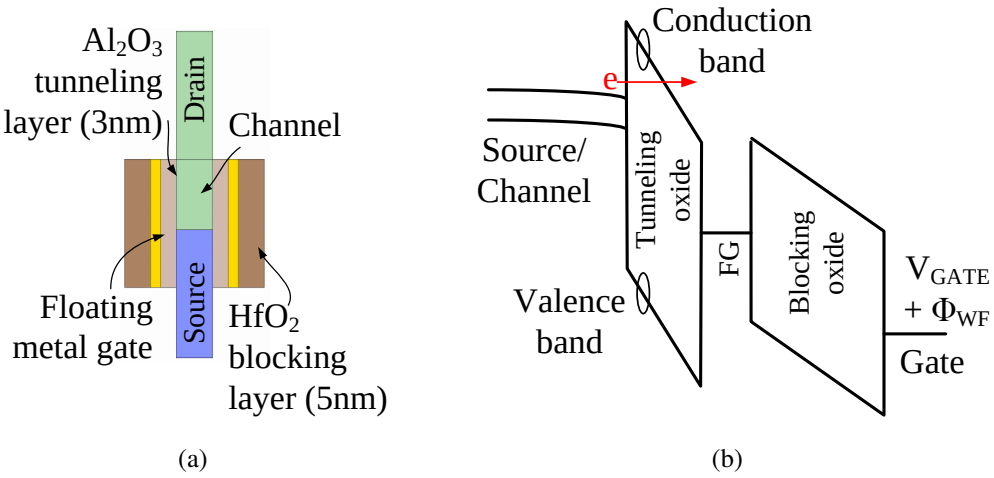
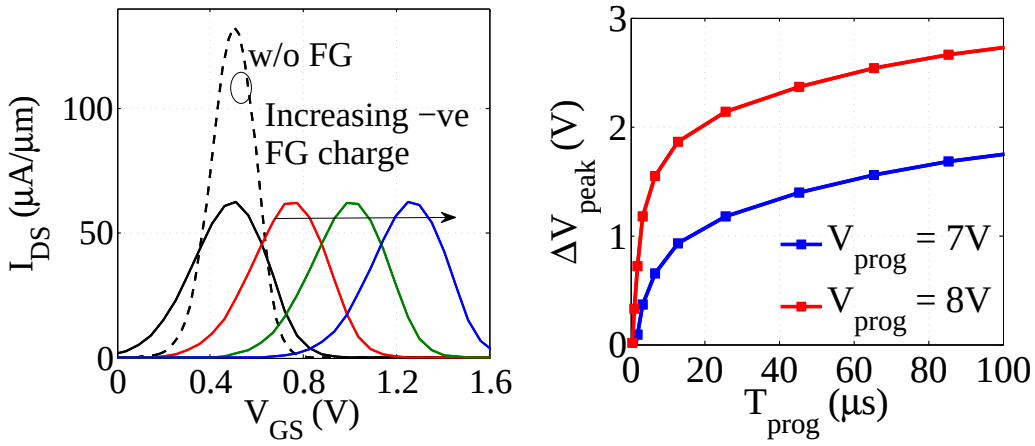



**Figure 62: Programming Gaussian-$I_{DS}$-$V_{GS}$ of FG-SO-HTFET: (a) FG-SO-HTFET schematic, and (b) charge-injection to the floating-gate at sufficiently high gate-programming-voltage.**

**Table 7: Direct-tunneling parameters for various materials**

| Effective mass | $GaAs_{0.1}Sb_{0.9}$ [99] | InAs [99] | $Al_2O_3$ [100] | $HfO_2$ [101] |
|---|---|---|---|---|
| $m_e/m_0$ | 0.043 | 0.023 | 0.20 | 0.08 |
| $m_h/m_0$ | 0.053 | 0.026 | 0.25 | 0.5 |

Programming of Gaussian-$I_{DS}$-$V_{GS}$ of FG-SO-HTFET is shown in Figure 63. With increasing negative-charge in the floating-gate, $V_{peak}$ of Gaussian-$I_{DS}$-$V_{GS}$ is programmed to a higher voltage [Figure 63a]. The peak sharpness is reduced in FG-SO-HTFET because of the thicker gate-dielectric-thickness. Effective charge injected to the floating-gate is modulated by varying programming-voltage ($V_{G,prog}$) and/or programming-duration ($T_{G,prog}$). $V_{peak}$ shift at varying $T_{G,prog}$ and $V_{G,prog}$ is shown in Figure 63b. With increasing $T_{G,prog}$, $V_{peak}$ shift saturates as the electric-field across the tunneling-oxide decreases.

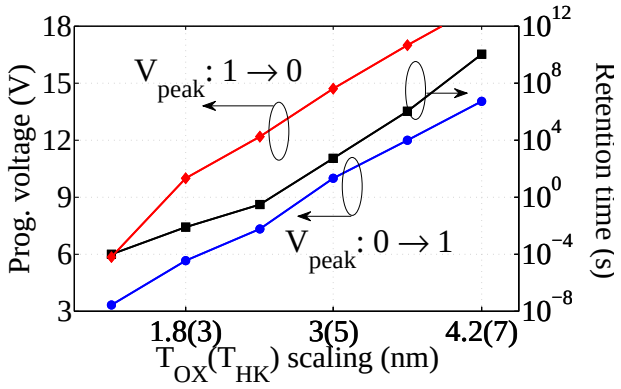



(a) (b)

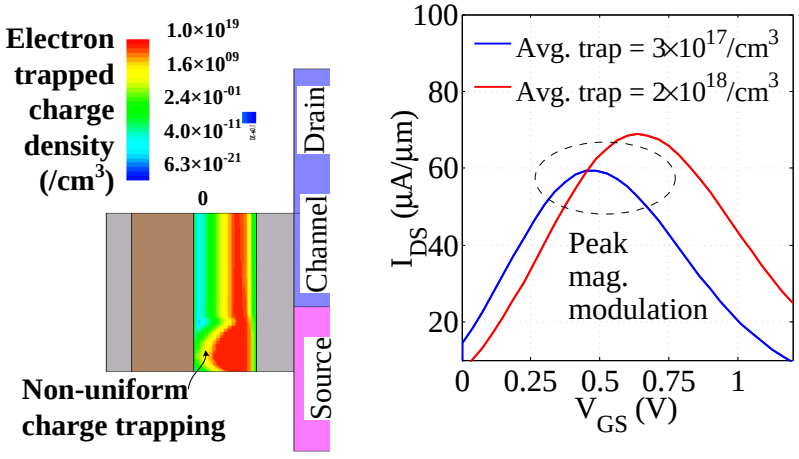
**Figure 63: $I_{DS}$-$V_{GS}$ characteristics of FG-SO-HTFET: (a) $V_{peak}$ programming by injecting charge to the floating-gate, and (b) $\Delta V_{peak}$ at varying programming-period ($T_{prog}$) and gate-programming-voltage ($V_{prog}$) [$V_{DS}$ = 0].**

Floating-gate thickness scaling in FG-SO-HTFET to reduce the programming-voltage

and/or duration is discussed in Figure 64. A lower floating-gate thickness can reduce $V_{prog}$, but it also reduces the charge-retention time [Figure 64a]. A nitride-trap-layer based SO-HTFET has a higher charge-retention than FG-SO-HTFET [102]. However, nitride-SO-HTFET also shows a non-uniform charge-trapping because of the heterogeneous source and channel materials [Figure 64b]. A non-uniform charge-trapping in nitride-SO-HTFET leads to different $I_{peak}$ at different $V_{peak}$ which inhibits AP [Figure 64c].




(a)




(b) (c)

**Figure 64: Design of the charge-trapping stack in FG-SO-HTFET: (a) programming-voltages and charge-retention at varying floating-gate thickness, (b) non-uniform charge-trapping in nitride layer-based SO-HTFET, and (c) modulation of both $V_{peak}$ and $I_{peak}$ in a nitride-SO-HTFET.**

## 8.4 SO-HTFET-based Associative Processing

SO-HTFET-based AP architecture is shown in Figure 65. In Figure 65a, source-terminals of SO-HTFETs are grounded. Drain-terminals of the SO-HTFETs within a column are shared. Gate-terminals of the SO-HTFETs within a row are shared. Each column in the array stores the 'feature vector' of a pattern (a $V_{peak}$ vector). The test-pattern (a $V_{test}$ vector) is applied simultaneously to all the columns. Drain-current through each SO-HTFET peaks when the applied test-pattern voltage is the same as its programmed-$V_{peak}$. If the applied test-pattern is different from the programmed-$V_{peak}$, drain-current through SO-HTFET is lower than the peak-current proportionally to the mismatch between the test-pattern and the programmed-$V_{peak}$. Therefore, the net current of all SO-HTFETs in a column follows the distance between $V_{peak}$ and $V_{test}$ [Figure 65b]. The column with the maximum-current, identified by a winner-take-all (WTA), represents the closest match (minimum $V_{test}$ to $V_{peak}$ distance). If all columns have lower current than $I_{th}$, a no match is determined. Note that the advantage of such non-Boolean-AP over the conventional designs [103] is the complete parallelism in operation.

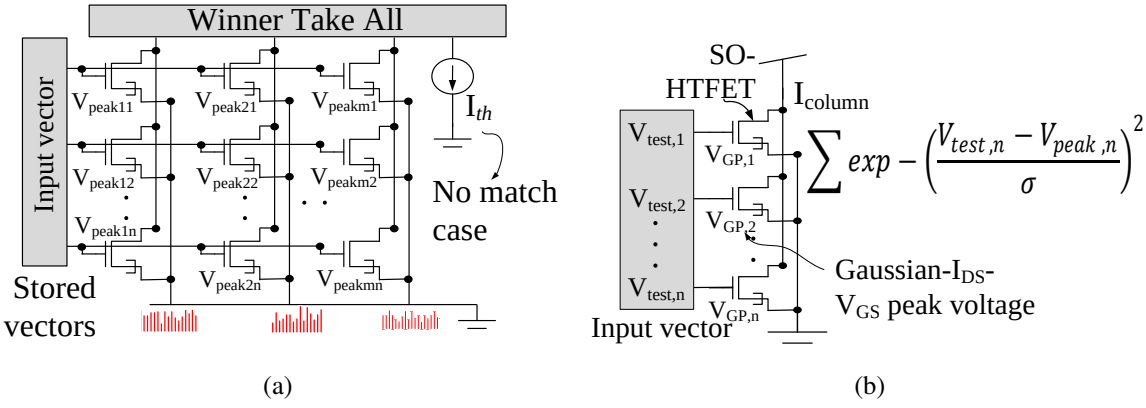
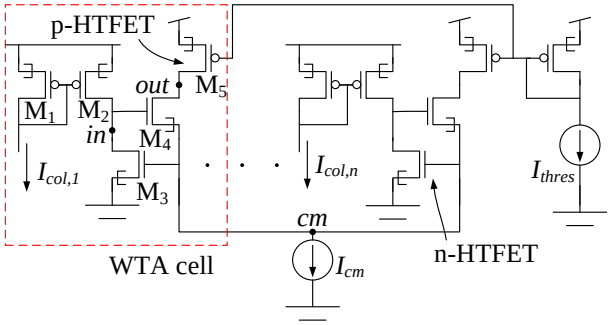


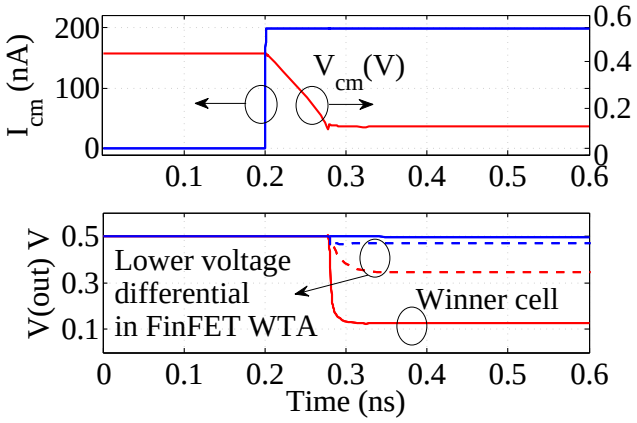

**Figure 65: SO-HTFET-based associative-processing: (a) architecture of SO-HTFET associative-processing array, and (b) column-current.**

## 8.5 HTFET Winner-Take-All

A winner-take-all (WTA) circuit based on [43] is shown in Figure 66. WTA identifies the column with the maximum current. WTA design is implemented with HTFETs. The HTFETs follow the geometrical/process specifications of Figure 58, however HTFETs are not implemented with a gate/source-overlap.




(a)




(b)

**Figure 66: SO-HTFET-AP peripheral: (a) HTFET winner-take-all, and (b) transients simulations.**

In Figure 66a, WTA is composed of the identical cells, where each cell senses the respective column-current ($I_{col}$) in the AP-array. The node 'cm' is shared among all the WTA-cells. The WTA bias-current ($I_{cm}$) is initialized to zero, and the node 'out' in each

WTA-cell is pre-charged to the supply-voltage (VDD). Current-mirror $M_1$-$M_2$ in a WTA-cell forces $I_{col}$ through the transistor $M_3$. The gate-voltage of $M_3$ is the same among all WTA-cells. Therefore, the potential of the node 'in' maximizes in the WTA-cell with the maximum $I_{col}$ (i.e., in the winning-cell), and $M_4$ in the winning-cell contributes maximally to the WTA bias-current, $I_{cm}$. The feedback action of $M_3$-$M_4$ in the winning-cell further starves $M_4$ in the other WTA-cells of $I_{cm}$, and eventually $I_{cm}$ flows solely through $M_4$ in the winning-cell. Hence, potential of the node 'out' drops in the winning-cell, while 'out' stays charged to VDD in the other WTA-cells.

Transient simulation waveforms for WTA is shown in Figure 66b. A digital latch reading the potential of the node 'out' in each WTA cell identifies the winner cell. Performance of HTFET-WTA is compared against FinFET-WTA, where FinFET-WTA is simulated using predictive technology model [104]. Notably, the steeper-switching-slope of HTFET than the conventional transistor, FinFET, leads to a larger output differential in WTA, and therefore, HTFET-WTA shows a more robust WTA operation.

## 8.6 On-line Training and Plasticity in SO-HTFET-AP

An on-line learning scheme for SO-HTFET-AP is shown in Figure 67. For each storage pattern, the AP-array in Figure 65a is required to store a 'feature vector' of the pattern in the respective column. An average of various samples of a pattern is used as the feature vector of the pattern [41]. The configuration in Figure 67a trains the SO-HTFET array to store the feature-vectors of the patterns by exploiting the floating-gate charge-plasticity. A test-pattern ($V_{test}$) is applied at the gates of the SO-HTFET array. The source (drain) of the storage column are driven successively with the programming voltages, $-V_{prog,P}$ and $V_{prog,N}$. In Figure 67b, $V_{prog,P}$ and $-V_{prog,N}$ are the threshold-voltages for FG-SO-HTFET $V_{peak}$ programming. In successive training pulses, gate-to-source $V_{prog}$ in SO-HTFET-DCC is $V_{test}+V_{prog,P}$ and $V_{test}-V_{prog,N}$. As a result of the training pulses, $V_{peak}$ of the SO-HTFET cells drift towards $V_{test}$ if $V_{peak}$ and $V_{test}$ differ [Figure 67c]. Under training with the several

samples of the pattern, such on-line training programs $V_{peak}$ in the respective column cells to an average of the samples (i.e., to the feature-vector of the pattern) [Figure 67d].

The on-line training can also be applied during operation to reinforce the queried pattern on the winning-column following the learning rule and implementation shown in Figure 67a. A 'reinforced learning' accrues the effect of even more sample patterns on the programmed $V_{peak}$ vector, and improves recognition accuracy.

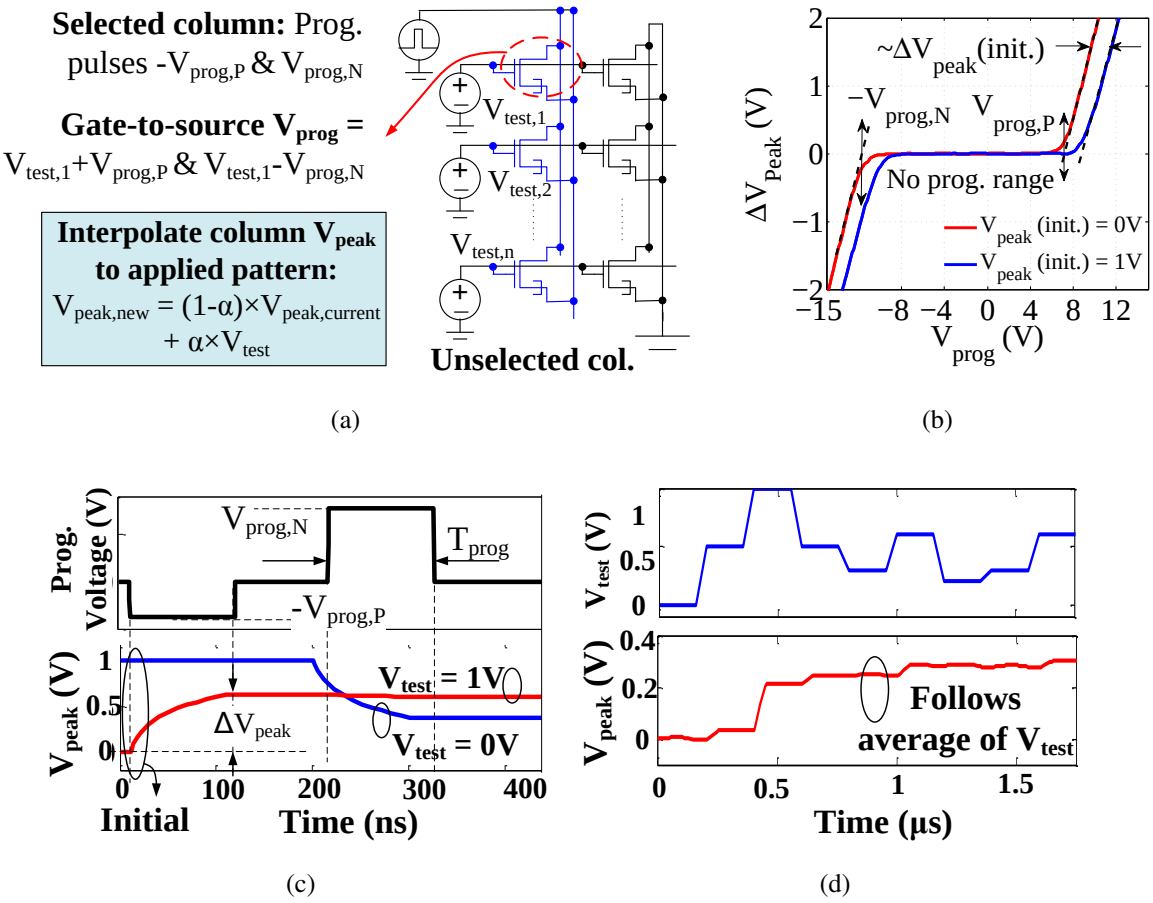
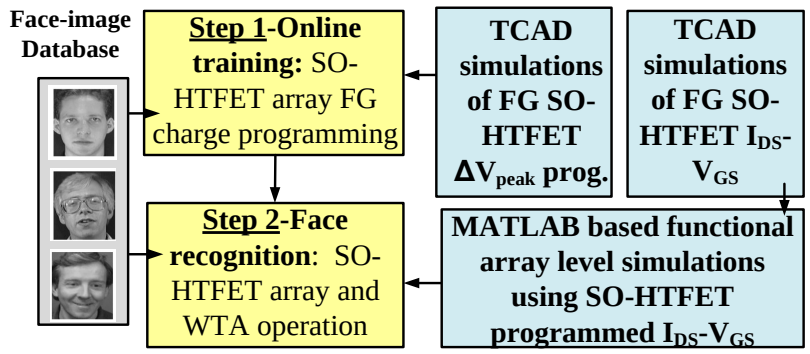


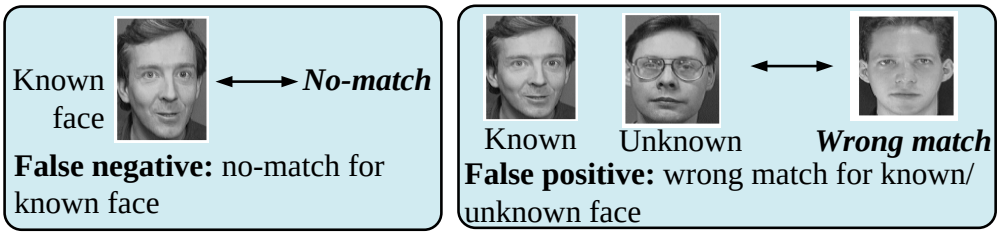

**Figure 67: On-line training in SO-HTFET-AP: (a) by column programming pulses and test-pattern $V_{test}$ at the array gates, (b) $V_{peak}$ modulation at varying program voltage ($V_{prog}$) [$T_{prog}$ = 100ns], (c) transients for initial $V_{peak}$ = 0 & 1V and $V_{test}$ = 1 & 0V, respectively, and (d) transients showing $V_{peak}$ to follow $V_{test}$ average over training steps.**

## 8.7 Face-recognition using SO-HTFET Associative Processing Platform

SO-HTFET-AP-based face-recognition is studied in this section [Figure 68]. A simulation methodology for face-recognition is shown in Figure 68a. Recognition-accuracy is estimated considering false-positives and false-negatives as shown in Figure 68b. Face-image database from the University of Cambridge [105] is used. The image database consists of face-images of 40 different persons, and 10 sample-images of each person. Statistical simulations over a thousand random runs extract face-recognition accuracy. In each sample run, the random image of a person from the database is selected as the test-pattern. The resolution of stored feature vector and the applied test image is reduced to the column length of SO-HTFET-AP array. For evaluation of both the false-negatives and false-positives in face-recognition, the SO-HTFET-AP array only stores 'feature vectors' of half of the face-images in the database. The voltage range of the applied pattern ($V_{test}$) and the stored vectors ($V_{peak}$) is $0-1.05$V.




(a)




(b)

**Figure 68: SO-HTFET-AP-based face-recognition: (a) the simulation methodology, and (b) considered false-positive and false-negative errors.**

Face-recognition simulation results are shown in Figure 69. For a realistic assessment of the recognition-accuracy under fabrication imperfections, process variability in SO-HTFET and imprecision in WTA is considered [Figure 69a]. A Gaussian distribution of variability is considered in SO-HTFET-$V_{peak}$ with $\sigma$=50mV and in SO-HTFET-$I_{peak}$ with $\sigma$=10%. An imprecision of 10% is considered in WTA.

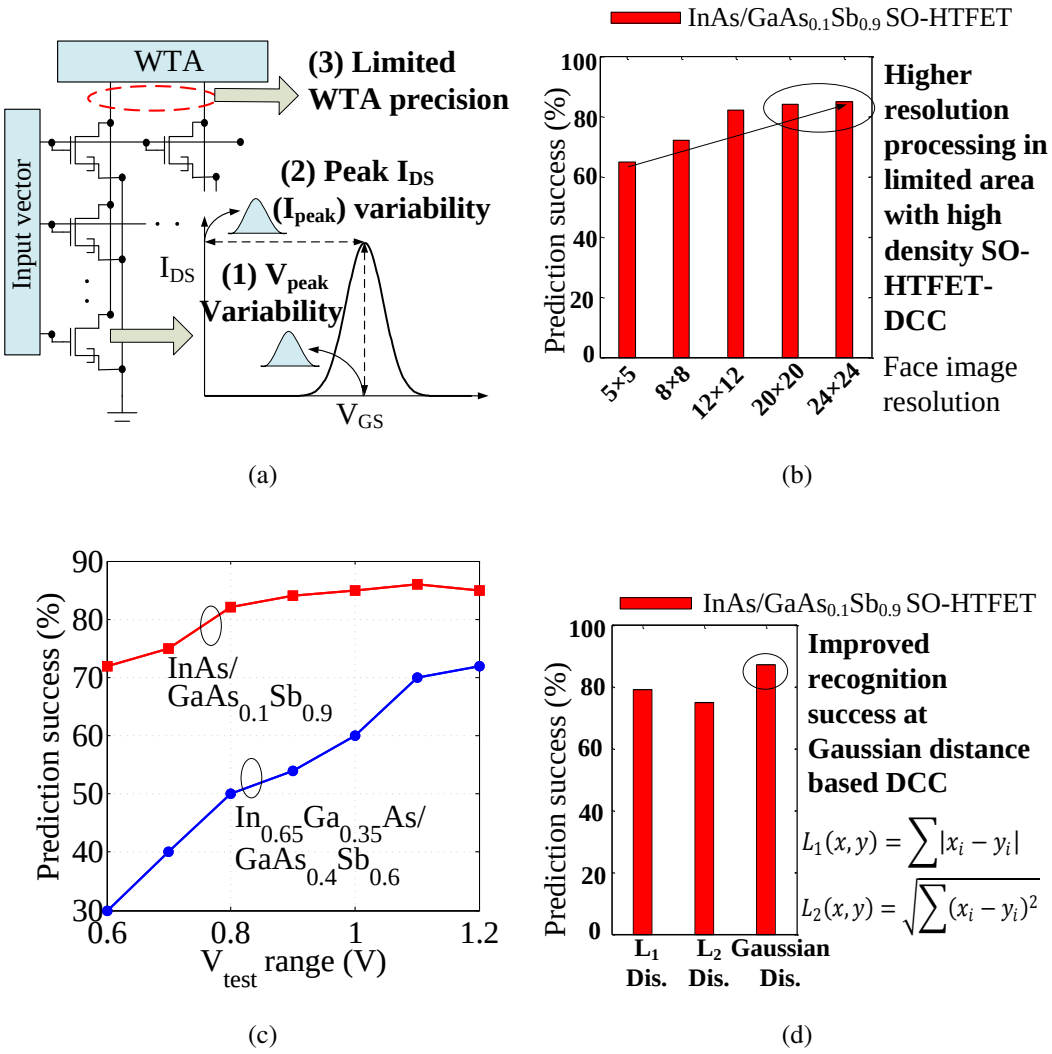



**Figure 69: Accuracy in SO-HTFET-AP-based face-recognition: (a) considered process imperfections in SO-HTFET and imprecision in WTA, (b) accuracy at varying resolution of AP, (c) accuracy in InAs/GaAs$_{0.1}$Sb$_{0.9}$-SO-HTFET and In$_{0.65}$Ga$_{0.35}$As/GaAs$_{0.4}$Sb$_{0.6}$-SO-HTFET-based AP across $V_{test}$ range, and (d) accuracy comparison between Gaussian distance-based AP (as in SO-HTFET-based AP) and L1/L2 norm-based AP (as in the conventional designs [3]).**

In Figure 69b, even considering SO-HTFET imperfections and WTA imprecision, the simulations show higher that 80% face-recognition accuracy. The accuracy is greater for a higher resolution images (i.e., in a longer column length AP-array), showing the need

for high-density DCC, as proposed in this chapter, to minimize the area of the platform. In Figure 69c, $In_{0.65}Ga_{0.35}As/GaAs_{0.4}Sb_{0.6}$ SO-HTFET-based AP has inferior accuracy than $InAs/GaAs_{0.1}Sb_{0.9}$ SO-HTFET-based AP because of the higher variance in the Gaussian-shape of $In_{0.65}Ga_{0.35}As$ /$GaAs_{0.4}Sb_{0.6}$ SO-HTFET. A higher variance in $In_{0.65}Ga_{0.35}As/GaAs_{0.4}Sb_{0.6}$ SO-HTFET reduces the wining column current to the loosing column current ratio, and makes the AP prone to WTA-imprecision. Notably, In Figure 69d, the Gaussian distance-based AP with SO-HTFET, as discussed in this chapter, also achieves higher accuracy than the conventional L1/L2 distance based AP [3].

## 8.8 Reinforced learning in SO-HTFET Associative Processing

A reinforced learning in SO-HTFET-AP is discussed in Figure 70. In Figure 70a, the reinforced learning assimilates query-patterns in the stored feature-vector of the predicted-pattern in the SO-HTFET array. Because of the feature-vector adaptation to the query-patterns, the recognition-accuracy with reinforced-learning improves. In Figure 70b, the reinforced-learning achieves 10% improvement compared to the without-reinforced-learning case. However, the learning factor ($\alpha$) in reinforced-learning should be optimized to represent a balance of the newly learned and past learned samples for maximum recognition-accuracy [Figure 70c].

Furthermore, while SO-HTFET-AP is susceptible to data decay due to floating-gate charge leakage, the reinforced learning can maintain a higher success rate by 'relearning' from the query patterns [Figure 71]. In Figure 71a, in a thin gate-stack (leakage time constant = 100ms) SO-HTFET array, the stored feature-vector decays because of floating-gate charge leakage. Hence, the prediction accuracy in SO-HTFET-AP without reinforced learning degrades over test-iterations in Figure 71b. However, a reinforced-learning interpolates the stored feature vector to query-patterns and maintains a higher recognition-accuracy.

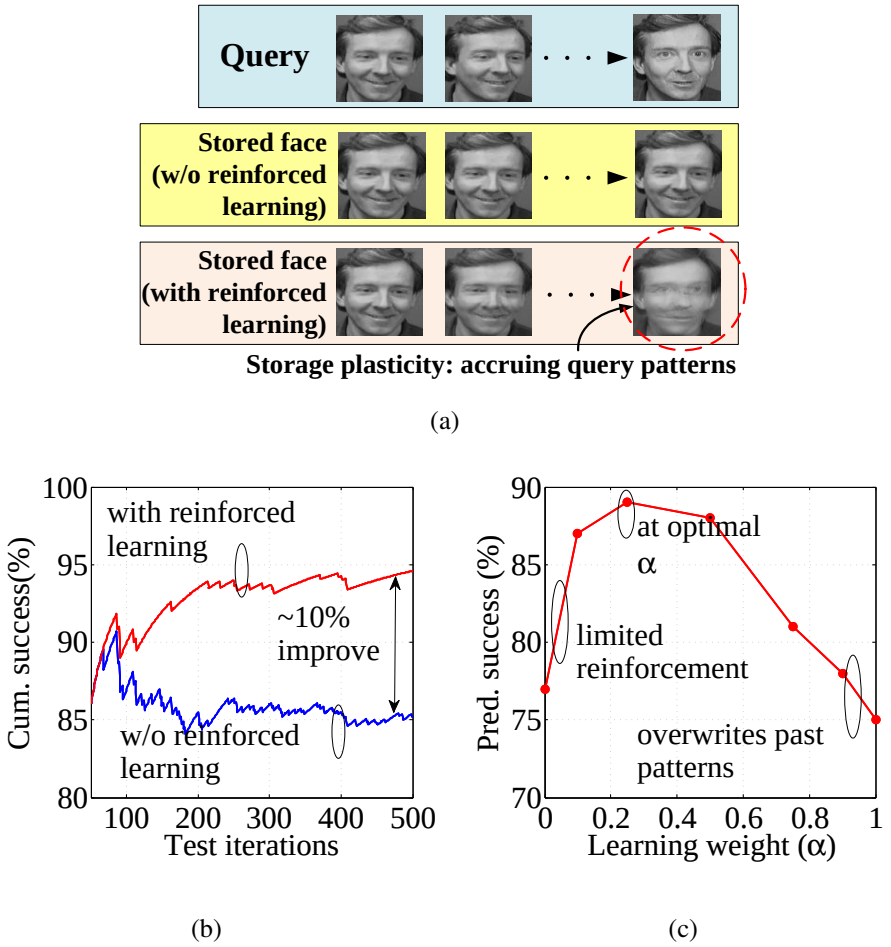



**Figure 70: Reinforced learning in SO-HTFET-AP-based face-recognition: (a) stored pattern plasticity demonstration, (b) improving prediction success with reinforced learning with test iterations, and (c) recognition-accuracy at varying learning-weight ($\alpha$).**

## 8.9 Comparison with the Existing Approaches

SO-HTFET-AP is compared against the existing platforms [Figure 72] in Table 8. In Table 8, SO-HTFET-DCC achieves smallest area among the existing techniques and only requires a single transistor for its implementation. Inter-cell coupling is also simplified in SO-HTFET-DCC, and only interconnects are required to implement such coupling by arranging DCCs in parallel. Because of its current-based output, SO-HTFET-DCCs can employ simplified, wide-range current based WTAs as peripherals. A complete parallelism in WTA-operation results in a higher performance in SO-HTFET-AP. SO-HTFET-AP also

features plasticity where the stored 'feature-vector' can be adapted run-time depending on the input-queries to improve recognition-accuracy. Because of its simplified architecture, SO-HTFET-AP has 250× lower power than digital-ASIC, and it achieves smallest DCC among the existing platforms.

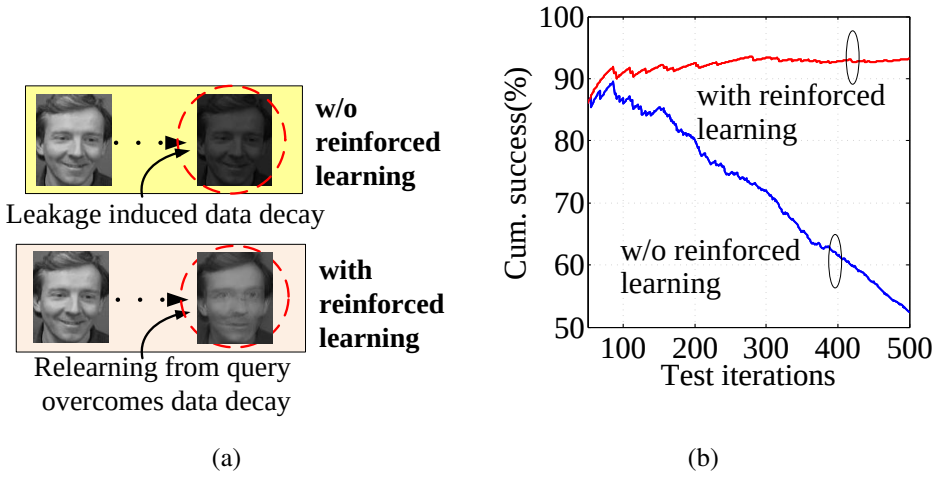



**Figure 71: Reinforced learning under floating-gate charge leakage: (a) floating-gate charge leakage resulting in stored pattern decay and (b) cumulative recognition accuracy with and without reinforced-learning in a thin-gate SO-HTFET-AP.**

## 8.10 Conclusions

SO-HTFET is attractive for non-Boolean AP because of its Gaussian-$I_{DS}$-$V_{GS}$ characteristics. Floating-gate configuration of SO-HTFET enables programming of Gaussian-$I_{DS}$-$V_{GS}$-$V_{peak}$. SO-HTFET-based AP compares the applied test-pattern simultaneously to all the stored patterns, and a complete parallelism of operation exhibits higher performance. As a result of majority-voting-based operation principle, SO-HTFET-based AP is also process variability resilient and attractive for technology-scaling. HTFET-WTA exploits steeper-switching-slope of HTFETs and shows a more robust operation than FinFET-WTA. An on-line training scheme, as discussed in the chapter, enables SO-HTFET-AP to on-line learn and store 'feature-vector' of a pattern. A reinforced-learning, as discussed in this

chapter, enables floating-gate charge-plasticity based on query-patterns and improves AP-accuracy. Reinforced-learning also overcomes limitation of charge-decay in the floating-gates, where the 'feature-vector' is relearned through query-patterns. SO-HTFET-based Gaussian-distance-AP improves recognition accuracy than equivalent L1/L2-norm-based AP. As compared to the existing AP-platforms, SO-HTFET-DCC facilitates area-efficient and energy-efficient AP.

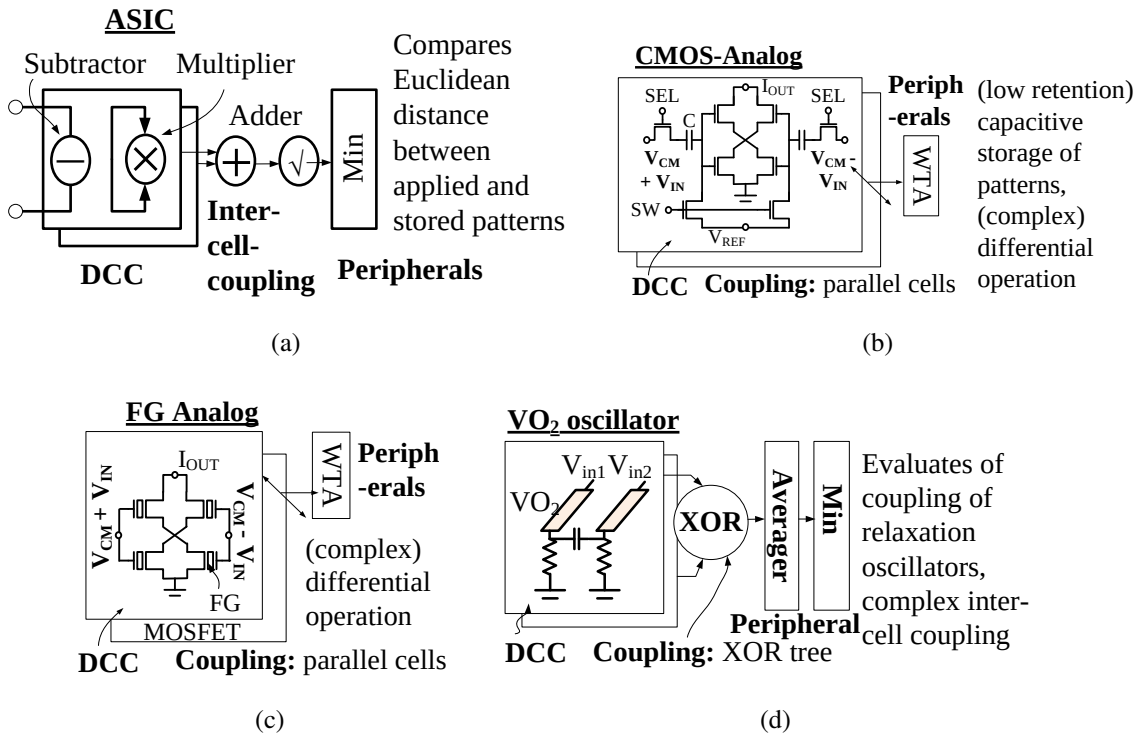



**Figure 72: Existing AP approaches: (a) digital ASIC, (b) CMOS-analog, (c) floating-gate CMOS-analog, and (d) $VO_2$ oscillators-based.**

**Table 8: Comparison of SO-HTFET-AP to the prior-works**

| | CMOS ASIC [3] | CMOS Analog [97] | FG CMOS | $VO_2$ Oscillator [3] | **SO-HTFET** |
|---|---|---|---|---|---|
| DCC implementation | Euclidean distance | Gaussian distance | Gaussian distance | Oscillators interlocking | Gaussian distance |
| DCC elements | >100 transistors | 8 transistors + 2 capacitors | 4 transistors | 2 transistors + 2 $VO_2$ + capacitor | **1 transistor** |
| DCC-coupling implementation | Adder tree based summation | Parallel DCCs | Parallel DCCs | XOR tree based averaging | Parallel DCCs |
| DCC-coupling elements | >50 transistors for 16 DCCs | Interconnects only | Interconnects only | >50 transistors for 16 DCCs + capacitor | **Interconnects only** |
| Peripheral implementation | Minimize sq. root | Winner-take-all | Winner-take-all | Minimize XOR average | Winner-take-all |
| Peripheral elements | Sq. root table + comparator | 5 transistors per column | 5 transistors per column | Counter + comparator | **5 transistors per column** |
| Plasticity | Write back to memory | No known technique | No known technique | No known technique | **Reinforced-learning** |
| AP power 16 (256) pixels/1GHz | 3mW(48mW) | 48$\mu$W(750$\mu$W) | 28$\mu$W(430$\mu$W) | 250$\mu$W(4mW) | **12$\mu$W(185$\mu$W)** |
| Comments | Based on 22nm FinFET design in [3] | Simulations with 22nm FinFET predictive model [104] | Simulations with 22nm FinFET predictive model [104] projected to floating-gate stack | Based on 22nm technology compatible design in [3] | **250× lower power than ASIC, smallest DCC** |

# CHAPTER 9

# SUMMARY AND FUTURE WORKS

This chapter summarizes the key contributions of this dissertation, and also identifies related future research goals. The key contributions are as following:

- **Energy-Efficiency and Reliability Limitations in CMOS-based Computing**: This dissertation has studied the limitations of CMOS-based Computing. As discussed in Chapter 2, a limited switching slope of CMOS constrains the energy-efficiency of CMOS-based digital as well as analog computing. Chapter 3 has illustrated further emerging challenges with nano-scaled CMOS process and non-traditional integration. A novel variability mechanism specific to high-$\kappa$/metal gate transistors was studied in Chapter 3, where due to deposition of metal-gate to high-$\kappa$ dielectric, resulting oxygen vacancies induce transistor variability. Study of transistor variability mechanism, such as oxygen vacancy, is critical in ensuring reliability of semiconductor devices and electronic circuits. Chapter 3 also showed that the transistor variability is not only limited to intrinsic fabrication imperfections, but it can also arise from novel transistor assembly techniques. Studying a three-dimensionally integrated SOI devices, Chapter 3 reported variability induced by neighboring vertical interconnects due to the electrical coupling.

- **Application of Neuronal Principles for Energy-Efficient CMOS-based Computing**: Fundamental energy-efficiency limitations in CMOS-based computing, as discussed in Chapter 2, and emerging limitations, as discussed in Chapter 3, urge for a fresh perspective to develop novel circuit techniques overcoming these limitations. In Chapter 4, a power-gating efficiency leaner was discussed inspired by the neuronal learning principles. The learner circuit is area and power-effective and facilitates adaptation against history and process/temperature conditions. A test-chip in IBM 130 nm process demonstrated the functionality of the learner circuit. On a broader

note, the learner circuit also suggests tremendous potential of a design methodology integrating conventional and neuromorphic computing principles, and inspires future research work in the direction.

- **Exploration of Neuromorphic Computing at Ultra Low Power**: Chapter 6 has explored co-design of TFET and large scale cellular neural network (CNN). Particularly, low ON/OFF current of Si channel TFET, while not promising for digital applications, is found suitable for large scale CNN. By collective computing, a large scale CNN mitigates on-current deficiencies of Si channel TFET. Low off-current and subthermal slope of Si channel TFET operates a large scale CNN under lower power. In Chapter 7 co-designing TFET and CNN-based associative memory, a higher neighborhood radius CNN exploits peculiar decreasing switching slope of TFET to enhance energy-efficiency of associative computing. Although, at lower power the variability of CNN computing elements aggravates; a higher neighborhood radius architecture averages and negates imperfections and mitigates variability challenges. Thereby, a co-design of TFET and CNN achieves higher energy efficiency and reliability even at lower power.

- **Exploration of the Technology for Integrated Conventional and Neuromorphic Computing**: Chapter 4 demonstrated the potential of a design approach integrating conventional and neuromorphic computing units. For an efficient, seamless, and low-cost integration of conventional and neuromorphic computing, it will be imperative to explore transistor technologies excelling in both the computing approaches. Chapter 6 & 7 have shown the potential of silicon channel SiGe-TFET in energy-efficient low-power neural networks. Note that SiGe-TFET is a compatible technology with the conventional CMOS, hence CMOS-based conventional computing units can be seamlessly integrated with SiGe-TFET-based neural networks. Chapter 6 has also

shown the potential of IIIV-TFET in high performance energy-efficient cellular neural networks. IIIV-TFETs are also attractive for energy-efficient digital computing due to its steeper switching slope. Therefore, IIIV-TFET can also be a technology of interest to integrate conventional and neuromorphic computing.

- **Rethinking Transistor Designs for Non-Boolean Computing**: Chapter 8 has illustrated the role of non-conventional transistor designs for non-Boolean computing. The conventional digital Boolean computing relies on the switch-like characteristics of a transistor, where a high ON-to-OFF current ratio becomes imperative for the energy-efficiency of the design. Meanwhile, Chapter 8 has shown that the exotic characteristics of non-conventional transistors can be capitalized by various non-Boolean platforms to enhance their energy-efficiency. Particularly, a source/gate overlap heterojunction TFET (SO-HTFET) design was demonstrated to exhibit a unique negative gate transconductance (NGT) characteristic. This NGT characteristic of TFET realized associative processing non-Boolean cell with a single transistor. Thereby, a co-designed computing device presented exciting opportunities of compact and high memory capacity non-Boolean pattern matching.

# REFERENCES


[1] A. Vandooren, D. Leonelli, R. Rooyackers, A. Hikavyy, K. Devriendt, M. Demand, R. Loo, G. Groeseneken, and C. Huyghebaert, "Analysis of trap-assisted tunneling in vertical si homo-junction and sige hetero-junction tunnel-fets," *Solid-State Electronics*, vol. 83, pp. 50–55, 2013.

[2] H. L. V. C. M. B. B. R. M. J. H. T. C. K. W. J.-H. K. D. G. K. P. C. J. S. R. E.-H. S. S. R. Pandey, H. Madan and S. Datta, "Demonstration of p-type in0.7ga0.3as/gaas0.35sb0.65 and n-type gaas0.4sb0.6/in0.65ga0.35as complimentary heterojunction vertical tunnel fets for ultra-low power logic," in *VLSI technology (VLSIT), 2015 symposium on*, IEEE, 2015.

[3] A. Parihar, N. Shukla, S. Datta, and A. Raychowdhury, "Synchronization of pairwise-coupled, identical, relaxation oscillators based on metal-insulator phase transition devices: A model study," *Journal of Applied Physics*, vol. 117, no. 5, p. 054902, 2015.

[4] R. Paradiso, G. Loriga, and N. Taccini, "A wearable health care system based on knitted integrated sensors," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 9, no. 3, pp. 337–344, 2005.

[5] L. Coetzee and J. Eksteen, "The internet of things-promise for the future? an introduction," in *IST-Africa Conference Proceedings, 2011*, pp. 1–9, IEEE, 2011.

[6] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature*, vol. 479, no. 7373, pp. 329–337, 2011.

[7] A. R. Trivedi, S. Carlo, and S. Mukhopadhyay, "Exploring tunnel-fet for ultra low power analog applications: a case study on operational transconductance amplifier," in *Design Automation Conference*, p. 109, ACM, 2013.

[8] Q. Huang, R. Huang, Z. Zhan, Y. Qiu, W. Jiang, C. Wu, and Y. Wang, "A novel si tunnel fet with 36mv/dec subthreshold slope based on junction depleted-modulation through striped gate configuration," in *IEEE International Electron Devices Meeting*, pp. 8–5, IEEE, 2012.

[9] S. Mookerjea, R. Krishnan, S. Datta, and V. Narayanan, "On enhanced miller capacitance effect in interband tunnel transistors," *IEEE Electron Device Letters*, vol. 30, no. 10, pp. 1102–1104, 2009.

[10] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano letters*, vol. 8, no. 2, 2008.

[11] P. Damle, T. Rakshit, M. Paulsson, and S. Datta, "Current-voltage characteristics of molecular conductors: two versus three terminal," *IEEE Transactions on Nanotechnology*, vol. 1, no. 3, pp. 145–153, 2002.

[12] K. Gopalakrishnan, P. B. Griffin, and J. D. Plummer, "Impact ionization mos (i-mos)-part i: device and circuit simulations," *IEEE Transactions on Electron Devices*, vol. 52, no. 1, pp. 69–76, 2005.

[13] C. W. Yeung, C. Shin, C. Hu, T.-J. K. Liu, *et al.*, "Feedback fet: A novel transistor exhibiting steep switching behavior at low bias voltages," in *International Electron Devices Meeting*, pp. 1–4, 2008.

[14] H. Kam, D. T. Lee, R. T. Howe, and T.-J. King, "A new nano-electro-mechanical field effect transistor (nemfet) design for low-power electronics," in *International Electron Devices Meeting*, pp. 463–466, IEEE, 2005.

[15] T. N. Theis and P. M. Solomon, "In quest of the next switch: prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2005–2014, 2010.

[16] N. Patel, A. Ramesha, and S. Mahapatra, "Drive current boosting of n-type tunnel fet with strained sige layer at source," *Microelectronics Journal*, vol. 39, no. 12, pp. 1671–1677, 2008.

[17] D. Mohata, R. Bijesh, Y. Zhu, M. Hudait, R. Southwick, Z. Chbili, D. Gundlach, J. Suehle, J. Fastenau, D. Loubychev, *et al.*, "Demonstration of improved heteroepitaxy, scaled gate stack and reduced interface states enabling heterojunction tunnel fets with high drive current and high on-off ratio," in *Symposium on VLSI technology*, pp. 53–54, IEEE, 2012.

[18] C. Anghel, H. Hraziia, A. Gupta, A. Amara, and A. Vladimirescu, "30-nm tunnel fet with improved performance and reduced ambipolar current," *IEEE Transactions on Electron Devices*, vol. 58, no. 6, pp. 1649–1654, 2011.

[19] M. Born, K. Bhuwalka, M. Schindler, U. Abelein, M. Schmidt, T. Sulima, and I. Eisele, "Tunnel fet: A cmos device for high temperature applications," in *International Conference on Microelectronics*, pp. 124–127, IEEE, 2006.

[20] K. Boucart, W. Riess, and A. M. Ionescu, "Lateral strain profile as key technology booster for all-silicon tunnel fets," *IEEE Electron Device Letters*, vol. 30, no. 6, pp. 656–658, 2009.

[21] K. K. Bhuwalka, J. Schulze, and I. Eisele, "Scaling the vertical tunnel fet with tunnel bandgap modulation and gate workfunction engineering," *IEEE Transactions on Electron Devices*, vol. 52, no. 5, pp. 909–917, 2005.

[22] L. Lattanzio, L. De Michielis, and A. M. Ionescu, "Electron-hole bilayer tunnel fet for steep subthreshold swing and improved on current," in *European Solid-State Device Research Conference*, pp. 259–262, IEEE, 2011.

[23] R. Gandhi, Z. Chen, N. Singh, K. Banerjee, and S. Lee, "Cmos-compatible vertical-silicon-nanowire gate-all-around p-type tunneling fets with-mv/decade subthreshold swing," *IEEE Electron Device Letters*, vol. 32, no. 11, pp. 1504–1506, 2011.

[24] A. L. Vallett, S. Minassian, P. Kaszuba, S. Datta, J. M. Redwing, and T. S. Mayer, "Fabrication and characterization of axially doped silicon nanowire tunnel field-effect transistors," *Nano letters*, vol. 10, no. 12, pp. 4813–4818, 2010.

[25] V. Nagavarapu, R. Jhaveri, and J. C. Woo, "The tunnel source (pnpn) n-mosfet: A novel high performance transistor," *IEEE Transactions on Electron Devices*, vol. 55, no. 4, pp. 1013–1019, 2008.

[26] S. H. Kim, S. Agarwal, Z. A. Jacobson, P. Matheu, C. Hu, and T.-J. K. Liu, "Tunnel field effect transistor with raised germanium source," *IEEE Electron Device Letters*, vol. 31, no. 10, pp. 1107–1109, 2010.

[27] A. Vandooren, D. Leonelli, R. Rooyackers, A. Hikavyy, K. Devriendt, M. Demand, R. Loo, G. Groeseneken, and C. Huyghebaert, "Analysis of trap-assisted tunneling in vertical si homo-junction and sige hetero-junction tunnel-fets," *Solid-State Electronics*, vol. 83, pp. 50–55, 2013.

[28] F. Mayer, C. Le Royer, J. Damlencourt, K. Romanjek, F. Andrieu, C. Tabone, B. Previtali, and S. Deleonibus, "Impact of soi, si 1-x ge x oi and geoi substrates on cmos compatible tunnel fet performance," in *International Electron Devices Meeting*, pp. 1–5, IEEE, 2008.

[29] G. Fiori and G. Iannaccone, "Ultralow-voltage bilayer graphene tunnel fet," *IEEE Electron Device Letters*, vol. 30, no. 10, pp. 1096–1098, 2009.

[30] V. Saripalli, A. Mishra, S. Datta, and V. Narayanan, "An energy-efficient heterogeneous cmp based on hybrid tfet-cmos cores," in *Design Automation Conference*, pp. 729–734, ACM, 2011.

[31] F. Conzatti, M. Pala, D. Esseni, E. Bano, and L. Selmi, "Strain-induced performance improvements in inas nanowire tunnel fets," *IEEE Transactions on Electron Devices*, vol. 59, no. 8, pp. 2085–2092, 2012.

[32] S. O. Koswatta, D. E. Nikonov, and M. S. Lundstrom, "Computational study of carbon nanotube pin tunnel fets," in *International Electron Devices Meeting*, pp. 518–521, IEEE, 2005.

[33] L. O. Chua and L. Yang, "Cellular neural networks: Applications," *IEEE Transactions on Circuits and Systems*, vol. 35, no. 10, pp. 1273–1290, 1988.

[34] M. T. Bohr *et al.*, "Interconnect scaling-the real limiter to high performance ulsi," in *International Electron Devices Meeting*, pp. 241–244, INSTITUTE OF ELECTRICAL & ELECTRONIC ENGINEERS, INC (IEEE), 1995.

[35] L. Chua and L. Yang, "Cellular neural networks: theory," *IEEE Transactions on Circuits and Systems*, vol. 35, pp. 1257–1272, Oct 1988.

[36] L. Wang, J. P. de Gyvez, and E. Sánchez-Sinencio, "Time multiplexed color image processing based on a cnn with cell-state outputs," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 6, no. 2, pp. 314–322, 1998.

[37] G. Grassi, "On discrete-time cellular neural networks for associative memories," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 48, no. 1, pp. 107–111, 2001.

[38] D. Liu and A. N. Michel, "Sparsely interconnected neural networks for associative memories with applications to cellular neural networks," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 41, no. 4, pp. 295–307, 1994.

[39] P. Szolgay, I. Szatmári, and K. László, "A fast fixed point learning method to implement associative memory on cnn's," *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART 1 FUNDAMENTAL THEORY AND APPLICATIONS*, vol. 44, pp. 362–365, 1997.

[40] A. C. Delbem, L. G. Correa, and L. Zhao, "Design of associative memories using cellular neural networks," *Neurocomputing*, vol. 72, no. 10, pp. 2180–2188, 2009.

[41] M. Sharad, D. Fan, and K. Roy, "Ultra low power associative computing with spin neurons and resistive crossbar memory," in *Proceedings of the 50th Annual Design Automation Conference*, p. 107, ACM, 2013.

[42] T. Yamanaka, T. Morie, M. Nagata, and A. Iwata, "A single-electron stochastic associative processing circuit robust to random background-charge effects and its structure using nanocrystal floating-gate transistors," *Nanotechnology*, vol. 11, no. 3, p. 154, 2000.

[43] J. Lazzaro, S. Ryckebusch, M. A. Mahowald, and C. A. Mead, "Winner-take-all networks of o (n) complexity," tech. rep., DTIC Document, 1988.

[44] A. R. Trivedi, T. Ando, A. Singhee, P. Kerber, E. Acar, D. J. Frank, and S. Mukhopadhyay, "A simulation study of oxygen vacancy-induced variability in/metal gated soi finfet," *Electron Devices, IEEE Transactions on*, vol. 61, no. 5, pp. 1262–1269, 2014.

[45] A. R. Trivedi and S. Mukhopadhyay, "Through-oxide-via-induced back-gate effect in 3-d integrated fdsoi devices," *Electron Device Letters, IEEE*, vol. 32, no. 8, pp. 1020–1022, 2011.

[46] J. Schaeffer, L. Fonseca, S. Samavedam, Y. Liang, P. Tobin, and B. White, "Contributions to the effective work function of platinum on hafnium dioxide," *Applied physics letters*, vol. 85, no. 10, 2004.

[47] S. Guha and P. Solomon, "Band bending and the thermochemistry of oxygen vacancies in ionic metal oxide thin films," *Applied Physics Letters*, vol. 92, no. 1, p. 012909, 2008.

[48] C. Kittel, *Introduction to solid state physics*. Wiley, 2005.

[49] D. Ceresoli and D. Vanderbilt, "Structural and dielectric properties of amorphous zro 2 and hfo 2," *Physical Review B*, vol. 74, no. 12, p. 125108, 2006.

[50] S. Guha and V. Narayanan, "Oxygen vacancies in high dielectric constant oxide-semiconductor films," *Physical review letters*, vol. 98, no. 19, p. 196101, 2007.

[51] P. Broqvist, A. Alkauskas, and A. Pasquarello, "Band alignments and defect levels in si-hfo2 gate stacks: Oxygen vacancy and fermi-level pinning," *Applied Physics Letters*, vol. 92, no. 13, p. 2911, 2008.

[52] H. Takeuchi, D. Ha, and T.-J. King, "Observation of bulk hfo2 defects by spectroscopic ellipsometry," *Journal of Vacuum Science & Technology A*, vol. 22, no. 4, pp. 1337–1341, 2004.

[53] E. Cartier, M. Hopstaken, and M. Copel, "Oxygen passivation of vacancy defects in metal-nitride gated hfo2/sio2/si devices," *Applied Physics Letters*, vol. 95, no. 4, p. 42901, 2009.

[54] 2012. [Online: http://http://www.synopsys.com/tools/tcad/Pages/default.aspx].

[55] K.-Y. Toh, P.-K. Ko, and R. G. Meyer, "An engineering model for short-channel mos devices," *Solid-State Circuits, IEEE Journal of*, vol. 23, no. 4, pp. 950–958, 1988.

[56] C. Lombardi, S. Manzini, A. Saporito, and M. Vanzi, "A physically based mobility model for numerical simulation of nonplanar devices," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 7, no. 11, pp. 1164–1171, 1988.

[57] X. Wang, A. R. Brown, B. Cheng, and A. Asenov, "Statistical variability and reliability in nanoscale finfets," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 5–4, IEEE, 2011.

[58] C. Chen, B. Wheeler, D. Yost, J. Knecht, C. Chen, and C. Keast, "Soi-enabled three-dimensional integrated-circuit technology," in *SOI Conference (SOI), 2010 IEEE International*, pp. 1–2, IEEE, 2010.

[59] J.-h. Sim and J. B. Kuo, "An analytical back-gate bias effect model for ultrathin soi cmos devices," *IEEE transactions on electron devices*, vol. 40, no. 4, pp. 755–765, 1993.

[60] A. R. Trivedi, W. Yueh, and S. Mukhopadhyay, "In situ power gating efficiency learner for fine-grained self-adaptive power gating," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 61, no. 5, pp. 344–348, 2014.

[61] A. R. Trivedi and S. Mukhopadhyay, "Self-adaptive power gating with test circuit for on-line characterization of energy inflection activity," in *2012 IEEE 30th VLSI Test Symposium (VTS)*.

[62] J.-J. Kim, J.-J. Kim, I.-J. Chang, and K. Roy, "Pvt-aware leakage reduction for on-die caches with improved read stability," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 1, pp. 170–178, 2006.

[63] K. Usami and N. Ohkubo, "A design approach for fine-grained run-time power gating using locally extracted sleep signals," in *Computer Design, 2006. ICCD 2006. International Conference on*, pp. 155–161, IEEE, 2007.

[64] S. Ishihara, M. Hariyama, and M. Kameyama, "A low-power fpga based on autonomous fine-grain power gating," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 8, pp. 1394–1406, 2011.

[65] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, and P. Bose, "Microarchitectural techniques for power gating of execution units," in *Proceedings of the 2004 international symposium on Low power electronics and design*, pp. 32–37, ACM, 2004.

[66] A. Youssef, M. Anis, and M. Elmasry, "A comparative study between static and dynamic sleep signal generation techniques for leakage tolerant designs," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 16, no. 9, pp. 1114–1126, 2008.

[67] K. Usami, T. Hashida, S. Koyama, T. Yamamoto, D. Ikebuchi, H. Amano, M. Namiki, M. Kondo, and H. Nakamura, "Adaptive power gating for function units in a microprocessor," in *Quality Electronic Design (ISQED), 2010 11th International Symposium on*, pp. 29–37, IEEE, 2010.

[68] S. Mukhopadhyay, S. Bhunia, and K. Roy, "Modeling and analysis of loading effect on leakage of nanoscaled bulk-cmos logic circuits," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 25, no. 8, pp. 1486–1495, 2006.

[69] K. Usami, Y. Goto, K. Matsunaga, S. Koyama, D. Ikebuchi, H. Amano, and H. Nakamura, "On-chip detection methodology for break-even time of power gated function units," in *Proceedings of the 17th IEEE/ACM international symposium on Low-power electronics and design*, pp. 241–246, IEEE Press, 2011.

[70] A. R. Trivedi, S. Carlo, and S. Mukhopadhyay, "Exploring tunnel-fet for ultra low power analog applications: a case study on operational transconductance amplifier," in *Proceedings of the 50th Annual Design Automation Conference*, p. 109, ACM, 2013.

[71] A. R. Trivedi, M. F. Amir, and S. Mukhopadhyay, "Ultra-low power electronics with si/ge tunnel fet," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pp. 1–6, IEEE, 2014.

[72] U. E. Avci, R. Rios, K. Kuhn, I. Young, *et al.*, "Comparison of performance, switching energy and process variations for the tfet and mosfet in logic," in *VLSI technology (VLSIT), 2011 symposium on*, pp. 124–125, IEEE, 2011.

[73] "http://www.synopsys.com/tools/tcad/pages/default.aspx," 2014.

[74] A. Schenk, "Rigorous theory and simplified model of the band-to-band tunneling in silicon," *Solid-State Electronics*, vol. 36, no. 1, pp. 19–34, 1993.

[75] G. Hurkx, D. Klaassen, and M. Knuvers, "A new recombination model for device simulation including tunneling," *Electron Devices, IEEE Transactions on*, vol. 39, no. 2, pp. 331–338, 1992.

[76] C.-T. Sah, R. Noyce, and W. Shockley, "Carrier generation and recombination in pn junctions and pn junction characteristics," *Proceedings of the IRE*, vol. 45, no. 9, pp. 1228–1243, 1957.

[77] K.-H. Kao, A. S. Verhulst, W. G. Vandenberghe, B. Soree, G. Groeseneken, and K. De Meyer, "Direct and indirect band-to-band tunneling in germanium-based tfets," *Electron Devices, IEEE Transactions on*, vol. 59, no. 2, pp. 292–301, 2012.

[78] O. Madelung, *Semiconductors: data handbook*. Springer Science & Business Media, 2012.

[79] G. Hellings, G. Eneman, R. Krom, B. De Jaeger, J. Mitard, A. De Keersgieter, T. Hoffmann, M. Meuris, and K. De Meyer, "Electrical tcad simulations of a germanium pmosfet technology," *Electron Devices, IEEE Transactions on*, vol. 57, no. 10, pp. 2539–2546, 2010.

[80] Y. Taur, T. H. Ning, *et al.*, *Fundamentals of modern VLSI devices*, vol. 2. Cambridge University Press Cambridge, 1998.

[81] W. G. Vandenberghe, A. S. Verhulst, G. Groeseneken, B. Soree, and W. Magnus, "Analytical model for a tunnel field-effect transistor," in *Mediterranean Electrotechnical Conference*, pp. 923–928, IEEE, 2008.

[82] M. G. Bardon, H. P. Neves, R. Puers, and C. Van Hoof, "Pseudo-two-dimensional model for double-gate tunnel fets considering the junctions depletion regions," *IEEE Transactions on Electron Devices*, vol. 57, no. 4, pp. 827–834, 2010.

[83] Á. Zarándy and C. Rekeczky, "Bi-i: a standalone ultra high speed cellular vision system," *Circuits and Systems Magazine, IEEE*, vol. 5, no. 2, pp. 36–45, 2005.

[84] S. Espejo, R. Dominguez-Castro, J. Huertas, E. Sánchez-Sinencio, *et al.*, "Smart-pixel cellular neural networks in analog current-mode cmos technology," *Solid-State Circuits, IEEE Journal of*, vol. 29, no. 8, pp. 895–905, 1994.

[85] M. Egmont-Petersen, D. de Ridder, and H. Handels, "Image processing with neural networksa review," *Pattern recognition*, vol. 35, no. 10, pp. 2279–2301, 2002.

[86] A. R. Trivedi and S. Mukhopadhyay, "Potential of ultralow-power cellular neural image processing with si/ge tunnel fet," *Nanotechnology, IEEE Transactions on*, vol. 13, no. 4, pp. 627–629, 2014.

[87] A. R. Trivedi, S. Datta, and S. Mukhopadhyay, "Application of silicon-germanium source tunnel-fet to enable ultralow power cellular neural network-based associative memory," *Electron Devices, IEEE Transactions on*, vol. 61, no. 11, pp. 3707–3715, 2014.

[88] M. Namba and Z. Zhang, "Cellular neural network for associative memory and its application to braille image recognition," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pp. 2409–2414, IEEE, 2006.

[89] B. Baird, M. W. Hirsch, and F. Eeckman, "A neural network associative memory for handwritten character recognition using multiple chua characters," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 40, no. 10, pp. 667–674, 1993.

[90] H. J. Mattausch, W. Imafuku, A. Kawabata, T. Ansari, M. Yasuda, and T. Koide, "Associative memory for nearest-hamming-distance search based on frequency mapping," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 6, pp. 1448–1459, 2012.

[91] C. M. Newman, "Memory capacity in neural network models: Rigorous lower bounds," *Neural Networks*, vol. 1, no. 3, pp. 223–238, 1988.

[92] P. Szolgay, I. Szatmári, and K. László, "A fast fixed point learning method to implement associative memory on cnn's," *IEEE Transactions on Circuits and Systems*, vol. 44, no. 4, pp. 362–365, 1997.

[93] A. Vandooren, D. Leonelli, R. Rooyackers, K. Arstila, G. Groeseneken, and C. Huyghebaert, "Impact of process and geometrical parameters on the electrical characteristics of vertical nanowire silicon n-tfets," *Solid-State Electronics*, vol. 72, pp. 82–87, 2012.

[94] A. R. Trivedi, K. Z. Ahmed, and S. Mukhopadhyay, "Negative gate transconductance in gate/source overlapped heterojunction tunnel fet and application to single transistor phase encoder," *Electron Device Letters, IEEE*, vol. 36, no. 2, pp. 201–203, 2015.

[95] H. L. S. D. A. R. Trivedi, R. Pandey and S. Mukhopadhyay, "Gate/source overlapped heterojunction tunnel fet for non-boolean associative processing with plasticity," in *IEEE Electron Device Meeting*, IEEE, 2015.

[96] A. Parihar, N. Shukla, S. Datta, and A. Raychowdhury, "Exploiting synchronization properties of correlated electron devices in a non-boolean computing fabric for template matching," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 4, no. 4, pp. 450–459, 2014.

[97] T. Shibata, "Computing based on the physics of nano devicesa beyond-cmos approach to human-like intelligent systems," *Solid-State Electronics*, vol. 53, no. 12, pp. 1227–1241, 2009.

[98] S. Datta, N. Shukla, M. Cotter, A. Parihar, and A. Raychowdhury, "Neuro inspired computing with coupled relaxation oscillators," in *Design Automation Conference*, pp. 1–6, ACM, 2014.

[99] Y. VulA, "Handbook series on semiconductor parameters, vol. 1, m. levinshtein, s. rumyantsev, m. shur, ed," 1996.

[100] A. Padovani, L. Larcher, V. Della Marca, P. Pavan, H. Park, and G. Bersuker, "Charge trapping in alumina and its impact on the operation of metal-alumina-nitride-oxide-silicon memories: Experiments and simulations," *Journal of Applied Physics*, vol. 110, no. 1, p. 014505, 2011.

[101] A. Campera, G. Iannaccone, and F. Crupi, "Modeling of tunnelling currents in hf-based gate stacks as a function of temperature and extraction of material parameters," *Electron Devices, IEEE Transactions on*, vol. 54, no. 1, pp. 83–89, 2007.

[102] T.-S. Chen, K.-H. Wu, H. Chung, and C.-H. Kao, "Performance improvement of sonos memory by bandgap engineering of charge-trapping layer," *Electron Device Letters, IEEE*, vol. 25, no. 4, pp. 205–207, 2004.

[103] C. R. Clark and D. E. Schimmel, "Scalable pattern matching for high speed networks," in *Field-Programmable Custom Computing Machines, 2004. FCCM 2004. 12th Annual IEEE Symposium on*, pp. 249–257, IEEE, 2004.

[104] "http://http://ptm.asu.edu/," 2014.

[105] "http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html," 2015.